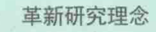


格致方法·定量研究系列　吴晓刚　主编

抽样调查方法简介

[美] 格雷汉姆·加尔顿（Graham Kalton）著
武玲蔚　译　周穆之　校



革新研究理念

丰富研究工具

最权威、最前沿的定量研究方法指南

40

格致出版社　上海人民出版社



格致方法·定量研究系列

1. 社会统计的数学基础
2. 理解回归假设
3. 虚拟变量回归
4. 多元回归中的交互作用
5. 回归诊断简介
6. 现代稳健回归方法
7. 固定效应回归模型
8. 用面板数据做因果分析
9. 多层次模型
10. 分位数回归模型
11. 空间回归模型
12. 删截、选择性样本及截断数据的回归模型
13. 应用logistic回归分析（第二版）
14. logit与probit：次序模型和多类别模型
15. 定序因变量的logistic回归模型
16. 对数线性模型
17. 流动表分析
18. 关联模型
19. 中介作用分析
20. 因子分析：统计方法与应用问题
21. 非递归因果模型
22. 评估不平等
23. 分析复杂调查数据（第二版）
24. 分析重复调查数据
25. 世代分析（第二版）
26. 纵贯研究（第二版）
27. 多元时间序列模型
28. 潜变量增长曲线模型
29. 缺失数据
30. 社会网络分析（第二版）
31. 广义线性模型导论
32. 基于行动者的模型
33. 基于布尔代数的比较法导论
34. 微分方程：一种建模方法
35. 模糊集合理论在社会科学中的应用
36. 图解代数：用系统方法进行数学建模
37. 项目功能差异（第二版）
38. Logistic回归入门
39. 解释概率模型
40. 抽样调查方法简介
41. 计算机辅助访问
42. 协方差结构模型：LISREL导论
43. 非参数回归
44. 广义线性模型：一种统一的方法
45. Logistic回归中的交互效应
46. 应用回归导论
47. 档案数据处理：研究“人生”
48. 创新扩散模型
49. 数据分析概论
50. 最大似然估计法：逻辑与实践

上架建议：社会研究方法

ISBN 978-7-5432-2419-

9 787543 224193 >

定价：22.00元

易文网：www.ewen.co

格致网：www.hibooks.cn

微信



微博

格致方法·定量研究系列　吴晓刚　主编

抽样调查方法简介

[美] 格雷汉姆·加尔顿（Graham Kalton） 著
武玲蔚 译　周穆之 校

SAGE Publications, Inc.

格致出版社　上海人民出版社

图书在版编目(CIP)数据

抽样调查方法简介/(美)加尔顿著;武玲蔚译;
周穆之校.—上海:格致出版社:上海人民出版社,
2014
(格致方法·定量研究系列)
ISBN 978-7-5432-2419-3

Ⅰ.①抽… Ⅱ.①加… ②武… ③周… Ⅲ.①抽样调
查统计-研究 Ⅳ.①C811

中国版本图书馆CIP数据核字(2014)第145554号

责任编辑 高 璇
美术编辑 路 静

格致方法·定量研究系列

抽样调查方法简介

[美]格雷汉姆·加尔顿 著
武玲蔚 译 周穆之 校

出 版 世纪出版股份有限公司 格致出版社 世纪出版集团 上海人民出版社 (200001 上海福建中路193号 www.ewen.cc)  编辑部热线 021-63914988 市场部热线 021-63914081 www.hibooks.cn	印 刷 浙江临安曙光印务有限公司 开 本 920×1168 1/32 印 张 5.5 字 数 106,000 版 次 2014年9月第1版 印 次 2014年9月第1次印刷
发 行 上海世纪出版股份有限公司发行中心	

ISBN 978-7-5432-2419-3/C·105 定价:22.00元

出版说明

由香港科技大学社会科学部吴晓刚教授主编的“格致方法 · 定量研究系列”丛书，精选了世界著名的 SAGE 出版社定量社会科学研究丛书，翻译成中文，起初集结成八册，于 2011 年出版。这套丛书自出版以来，受到广大读者特别是年轻一代社会科学工作者的热烈欢迎。为了给广大读者提供更多的方便和选择，该丛书经过修订和校正，于 2012 年以单行本的形式再次出版发行，共 37 本。我们衷心感谢广大读者的支持和建议。

随着与 SAGE 出版社合作的进一步深化，我们又从丛书中精选了三十多个品种，译成中文，以飨读者。丛书新增品种涵盖了更多的定量研究方法。我们希望本丛书单行本的继续出版能为推动国内社会科学定量研究的教学和研究作出一点贡献。

总 序

2003年，我赴港工作，在香港科技大学社会科学部教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课"社会科学里的统计学"（Statistics for Social Science）为所有研究型硕士生和博士生的必修课，而第二门课"社会科学中的定量分析"为博士生的必修课（事实上，大部分硕士生在修完第一门课后都会继续选修第二门课）。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的"消费者"而非"生产者"。作为"消费者"，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少

量重复,但各有侧重。“社会科学里的统计学”从介绍最基本的社会研究方法论和统计学原理开始,到多元线性回归模型结束,内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型,以及线性回归模型的假设和模型诊断。“社会科学中的定量分析”则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法,将重点放在因变量为定类数据的分析模型上,包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型,包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生,一直鼓励和支持我将两门课的讲稿结集出版,并帮助我将原来的英文课程讲稿译成了中文。但是,由于种种原因,这两本书拖了多年还没有完成。世界著名的出版社 SAGE 的“定量社会科学研究”丛书闻名遐迩,每本书都写得通俗易懂,与我的教学理念是相通的。当格致出版社向我提出从这套丛书中精选一批翻译,以飨中文读者时,我非常支持这个想法,因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种语言的精准把握能力,还要有对实质内容有较深的理解能力,而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容,只有语言能力是远远不能胜任的。在短短的一年时间里,我们组织了来自中国内地及香港、台湾地区的二十几位

研究生参与了这项工程,他们当时大部分是香港科技大学的硕士和博士研究生,受过严格的社会科学统计方法的训练,也有来自美国等地对定量研究感兴趣的博士研究生。他们是香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛,硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊,应用社会经济研究中心研究员李俊秀;香港大学教育学院博士研究生洪岩璧;北京大学社会学系博士研究生李丁、赵亮员;中国人民大学人口学系讲师巫锡炜;中国台湾"中央"研究院社会学所助理研究员林宗弘;南京师范大学心理学系副教授陈陈;美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛;美国加州大学洛杉矶分校社会学系博士研究生宋曦;哈佛大学社会学系博士研究生郭茂灿和周韵。

参与这项工作的许多译者目前都已经毕业,大多成为中国内地以及香港、台湾等地区高校和研究机构定量社会科学方法教学和研究的骨干。不少译者反映,翻译工作本身也是他们学习相关定量方法的有效途径。鉴于此,当格致出版社和 SAGE 出版社决定在"格致方法 · 定量研究系列"丛书中推出另外一批新品种时,香港科技大学社会科学部的研究生仍然是主要力量。特别值得一提的是,香港科技大学应用社会经济研究中心与上海大学社会学院自 2012 年夏季开始,在上海(夏季)和广州南沙(冬季)联合举办"应用社会科学研究方法研修班",至今已经成功举办三届。研修课程设计体现"化整为零、循序渐进、中文教学、学以致用"的方针,吸引了一大批有志于从事定量社会科学研究的博士生和青年学者。他们中的不少人也参与了翻译和校对的工作。他们在

繁忙的学习和研究之余，历经近两年的时间，完成了三十多本新书的翻译任务，使得“格致方法·定量研究系列”丛书更加丰富和完善。他们是：东南大学社会学系副教授洪岩璧，香港科技大学社会科学部博士研究生贺光烨、李忠路、王佳、王彦蓉、许多多，硕士研究生范新光、缪佳、武玲蔚、臧晓露、曾东林，原硕士研究生李兰，密歇根大学社会学系博士研究生王骁，纽约大学社会学系博士研究生温芳琪，牛津大学社会学系研究生周穆之，上海大学社会学院博士研究生陈伟等。

陈伟、范新光、贺光烨、洪岩璧、李忠路、缪佳、王佳、武玲蔚、许多多、曾东林、周穆之，以及香港科技大学社会科学部硕士研究生陈佳莹，上海大学社会学院硕士研究生梁海祥还协助主编做了大量的审校工作。格致出版社编辑高璇不遗余力地推动本丛书的继续出版，并且在这个过程中表现出极大的耐心和高度的专业精神。对他们付出的劳动，我在此致以诚挚的谢意。当然，每本书因本身内容和译者的行文风格有所差异，校对未免挂一漏万，术语的标准译法方面还有很大的改进空间。我们欢迎广大读者提出建设性的批评和建议，以便再版时修订。

我们希望本丛书的持续出版，能为进一步提升国内社会科学定量教学和研究水平作出一点贡献。

吴晓刚

于香港九龙清水湾

序

无论在学术界还是业界，抽样调查研究都是一个重要的领域。在社会科学的各个领域中，它都是很基础的工具，在很多大型调查中扮演了重要角色，包括全国选举调查(National Election Studies)、芝加哥大学的国情调查中心(National Opinion Research Center，NORC)的综合社会调查(General Social Survey)，以及密歇根大学调查研究中心(Survey Research Center)对消费者的调查等。这一方法几乎可以被用到任何情境中，包括描述性研究以及评估性研究。最后，抽样调查方法在政治竞选中的应用是非常成功的，从而极大地提高了自身的知名度。

当然，调查研究的基础是抽样过程。如果离开了设计优良、执行有力的抽样过程，无论研究者提出的研究问题多么有趣、使用的研究方法多么高端，都不能弥补这一缺憾。但即使这样，有人可能会觉得抽样仅仅是一个技术问题，只要统计学家懂就够了。我并不这么认为。负责抽样的统计学家对很多项目而言是至关重要的，使用抽样调查数据的研究者必须有足够的抽样调查的理论基础。

加尔顿教授的著作对抽样过程的介绍深浅适中。显然，本书并不是为统计学家而写的。事实上，本书对于那些只有部分统计学知识的研究者而言都是通俗易懂的。在书中，作者对所有的概念都进行了仔细解释，从而为读者理解调查设计打下了很好的基础。本书的一个主要特点在于，它对这一方法的实际应用进行了诸多介绍，比如抽样框和无应答的部分，而研究者在实际中常常会遇到无应答的问题。

本书很好地涵盖了抽样理论与相应实例，因此我认为，加尔顿教授的这一著作无论对于抽样调查的初学者还是有一定基础的读者来说，都是宝贵的参考资料。

理查德·G.涅米

目录

序 1

第 1 章　简介 1

第 2 章　简单随机抽样 7

第 3 章　系统抽样 19

第 4 章　分层抽样 25

第 1 节　按比例分层 29

第 2 节　非比例分层 33

第 3 节　层的选择 36

第 5 章　整群抽样和多阶抽样 39

第 6 章　按规模大小成比例的概率抽样 51

第 7 章 其他概率抽样设计 65
第 1 节 二阶段抽样 67
第 2 节 重复抽样 70
第 3 节 面板设计 76

第 8 章 抽样框 81
第 1 节 缺失元素 83
第 2 节 群 85
第 3 节 空白与外来元素 89
第 4 节 重复列举 91

第 9 章 无应答 93

第 10 章 调查分析 103
第 1 节 权重 105
第 2 节 抽样误差 113

第 11 章 样本量 123

第 12 章 两个例子 129
第 1 节 全国性面访调查 131
第 2 节 电话访问调查的例子 134

第 13 章 非概率抽样 139

第 14 章 结语 145

参考文献 148

译名对照表 151

第1章

简　介

目前，抽样调查（sample surveys）作为一种提供统计数据的方式，已经在众多领域被研究者和管理者们广泛应用。这些领域包括社会学、社会心理学、人口学、政治学、经济学、教育学以及公共健康等领域，人们使用抽样调查来发展、检验以及提炼他们的研究假设。与此同时，中央政府也同样依赖这些调查来获得关于民众的信息，包括就业和失业、收入和支出、住房条件、教育、营养、健康、出行方式，等等。他们也会对例如制造商、零售商、农场、学校以及医院等机构进行调查。另外，地方政府也会使用调查来帮助他们的规划。市场调查员也使用抽样调查来确定目标市场，了解商品是如何在实际中被使用的以及消费者的反响。意见调查则追踪政治家或政党的受欢迎程度，同时度量公众在众多社会议题上的看法。

虽然抽样调查现在有非常广泛的应用，但令人惊讶的是它的历史非常短。这一历史基本发生在20世纪之内，而调查方法的改进也大多出现在20世纪30年代之后。在这一世纪中，调查方法的所有方面都有了可观的改进，抽样方法方面的进步尤甚，而后者也正是本书的主题。20世纪之初，统计学家就仅仅分析总体的一部分是否足够进行了争论，因

为这在原理上是可行的(O'Muircheartaigh and Wong, 1981)。由于时间抽样(time sampling)已经被广泛接受,因此人们已经能够使用众多抽样方法在不同情境下来保证调查的有效性和实用性。

调查的设计包含了众多互相关联的决策,比如收集数据的方式(面对面的访问、电话访问或自我完成的形式)、提问框架、数据处理方法以及样本设计(Moser and Kalton, 1971; Warwick and Lininger, 1975)。尽管这本书仅涉及样本的设计,但它同时也会考虑到样本设计需要的是整个抽样调查的有机整体。特别是数据搜集过程中包含的经济学对样本设计的选择具有重要影响。

调查设计的第一步就是决定研究的总体(population)。这里,"总体"这一术语的意思是被研究的元素(element)的全体,而元素则是研究的单元。具体而言,元素可以是个人,也可以是家庭、农场、学校或其他单位。根据调查的对象,需要精确和仔细地定义"总体"的概念,因为研究的结果取决于我们使用的定义。比如,考虑一个在城市中进行的调查,其目标是检验一个对新引入的巴士系统的支持程度。我们是否应将调查设定到在这一城市里居住的个人层面?访问对象的最低年龄是什么?是否应当调查那些没有城市选举权的人?在城市中暂居的访客是否应当被排除在外?如果是的话,这一群体该如何定义?在我们定义总体时,会面临很多类似的问题,而这一工作并不像看起来那么简单。

在开始阶段,定义一个满足调查目标的理想总体是有帮助的:目标总体(target population)。这一定义现在常常被修

正为调查总体(survey population),从而将实际的限制纳入考虑范围内。比如,美国很多全国性的调查的理想状态是包含驻扎在海外的军人、居住在夏威夷或者阿拉斯加的人们,以及在医院、旅馆、监狱、军营和其他机构的人们。然而,对这些人们进行调查访问无疑会面临很多问题,在现实中这些人往往会被排除在调查总体之外。因此,从理想目标总体开始的优势在于,人们可以清楚地确定需要排除哪些人,从而使人们可以评估这些限制条件的范围和后果。

一旦定义了总体,我们就可以从中确定样本。一个最直接的方法就是将总体中的所有元素都包括进来,但是这通常并不合适。只从总体中的部分搜集信息的成本较低,同时如果可以保证我们之后的估计量足够准确的话,抽样明显是更加经济的做法。这样的做法也可以使这一过程更加迅速,人们也可以从中得到及时的报告。另外,通过仅关注总体中的一部分信息,数据搜集的质量会高于收集总体全部的做法。因此,抽样调查实际上能够提供更加准确的结果。因为这些理由,除非总体本身非常小,抽样调查几乎总是更常被使用的方法。

在样本设计时,人们往往考虑如何选择总体的部分来对其进行调查。一个基本的区分就是要看抽样是不是通过概率机制(probability mechanism)实现的。对于一个概率样本,每一个元素都有一个已知的、非零的被抽中的概率。因此,人们可以避免选择偏差,并且通过统计理论来推导出抽样估计量(survey estimator)的性质。非概率抽样则包含了多样的过程,包括使用志愿者以及特意选择某些具有"代表性"的元素作为样本。所有非概率抽样的弱点在于其主观

性，从而排除了人们为此发展出一个相应的理论框架的可能性。一个专家选择的志愿者的样本或者一个代表性样本仅仅能够被主观地评估，因而不能用不依赖这些主观性的统计方法来评判。考虑到非概率抽样的这一弱点，本书将仅仅考虑概率抽样。然而，我们在第 13 章中仍然会对非概率抽样作出一些讨论。

任何形式的概率抽样的基础都是抽样框（sampling frame），从中人们可以决定抽取哪些元素。在一个简单的情形下，当包含样本中的所有元素的列表存在时，这一列表就是抽样框。当我们没有列表时，抽样框就相当于一个确认总体中的元素的等价程序。地区抽样（area sampling）就是一个相应的例子。在这一技术下，总体中的每一个元素都与一个特定的地理位置相关联（比如，居民或者住户总有他们的居住地址；当居住地多于一个地址时，我们考虑其主要住址）。因此，当人们绘出一个地区的样本之后，这些被选择的地区中的所有元素或者是部分元素都会被纳入样本（见第 12 章）。抽样框的一般组织和它包含的元素的信息对于样本设计选择的影响常常很大。抽样框中的缺陷，比如如果不能将所有元素包含到其抽样总体中，可能就会对样本的选取有负面影响。我们会在第 8 章中对抽样框及其细节做更细致的探讨。

现在，人们已经发展出各种各样的概率抽样技术，它们能够提供有效的实际样本设计。其中，被广泛使用的是系统抽样（systematic sampling）、分层抽样（stratification）、多阶段抽样（multistage）、群（cluster），以及按规模大小成比例的概率抽样（probability proportional to size sampling）。以下内容

将会对这些方法分别进行探讨和解释，但是实际上它们常常被联合使用于一些复杂的样本设计中。第 12 章中的两个例子会说明这一点。下面，我们将从比较简单的、适合从比较紧凑的总体中抽取小样本的抽样方法开始，然后介绍适合从更大的、更分散的总体中抽样的复杂方法。

第2章

简单随机抽样

简单随机抽样(Simple Random Sampling, SRS)给人们提供了一个讨论概率抽样方法的自然的出发点,可并不是因为它的广泛使用(实际上确实是并不广泛),而是因为它是最简单的方法,并且是更加复杂的方法的基础。在定义简单随机抽样方法之前,我们先将样本量记为 n,将总体中的所有元素的个数记为 N。正式定义的简单随机抽样是使得任何包含 n 个元素的集合在总体的 N 个元素中具有相同抽取概率的抽样方法。这一定义表明,总体中的任一元素都有相同的概率被抽中,但上面的定义比这一描述更严格。下面我们将会看到,更加复杂的抽样方法也往往是等概率抽样(Equal Probability Selection Methods, EPSEM),但这些方法下的被抽取元素的集合的联合概率并不像简单随机抽样一样是相等的。

下面,我们会讨论简单随机抽样的一个特定的应用。假设人们将在一所高中进行调查以了解学生们的业余爱好。我们有这所学校的 1 872 名学生的名单,其中学生按照他们的身份码排列。这些身份码的范围是从 0001 到 1917,其中的一些间断是由于一些学生离开了学校。假设我们考虑使用简单随机抽样抽取一个 $n=250$ 的样本(在第 11 章中,我

们会讨论 n 的选取)。

一个按照简单随机抽样抽取的方法是使用抽奖方法(lottery method)。每个学生的名字或者其身份码被写到1 872个相同的圆片上。这些圆片完全打散放到一个罐子里,人们从中随机选择250个。如果这些程序得到完美的执行,那么这250个圆片就会确定简单随机抽样抽取的250个学生。尽管这一过程看着简单,但是做起来却很繁杂,因为它必须依赖于人们将这些圆片完全打散,从而保证随后的抽取是随机的。

另一个使用简单随机抽样抽取的方法是通过随机数表(table of random numbers)。这些表格是人们精心创建并检验的,以保证从长期来看每一数位、数位的每一组合等都是以相同的频率出现的。在表2.1中,我们给出了一个肯德尔和史密斯(Kendall and Smith, 1939)创建的随机数表的一部分。

表2.1 随机抽样数

67	28	96	25	68	36	24	72	03	85	49	24
85	86	94	78	32	59	51	82	86	43	73	84
40	10	60	09	05	88	78	44	63	13	58	25
94	55	89	48	90	80	77	80	26	89	87	44
11	63	77	77	23	20	33	62	62	19	29	03

资料来源:Kendall, M.G. and B.B.Smith, *Tables of Random Sampling Numbers*. Copyright © 1939 by Cambridge University Press. Reprinted by Permission.

由于每个学生的身份码中都包含四位数字,我们需要选择包含四位数字的随机数。在实际运用中,人们应当从表中的任意一处开始选择,但这里为了简单起见,我们将从左上

角开始。接着,我们先从首四列开始向下选取,然后是后面四列向下选取,依次进行。落在学生身份码范围之外的数字(0001—1917)或者在此范围内但查无此人的数字会被忽略。表 2.1 中的第一组的前四个四位数字(6728, 8586, 4010, 9455)没有产生任何可行的学生码,因此选择的第一个学生号码是 1163(并且此学生仍在学校)。继续看表格,另外两名被选中的学生是 0588 和 0385。显然,根据这一表格选取 250 名学生是一项索然无味的工作,这要求人们抽取出大量的随机数,而其中的大多数却并不能产生有效的学生码。

为了避免这些随机数的浪费,我们可以为每个学生指定多于一个的随机数,但前提是每个学生对应的随机数数量是相等的。这里,每个学生都可以与五个四位的随机数相联系。对于学生 0001 而言,一个简单的方法是让他与 2001, 4001, 6001 和 8001 相连;学生 0002 与 2002, 4002, 6002, 8002 相连;对于学生 1917,则让他与 3917, 5917, 7917 和 9917 相连。然后,我们再次从表 2.1 的左上角开始,被选择的学生是 6728＝学生 0728, 8586＝学生 0586, 4010＝学生 0010, 9455＝学生 1455, 1163＝学生 1163,等等。

使用随机数表来抽取样本可能会使得抽取同一元素的次数超过一次。而对于上面的抽彩方法而言,并不存在这一可能性,因为当一个学生的圆片被抽到时,我们并不将其放回罐子中去。然而,如果我们在下一次抽取前将圆片放回罐子中,这种可能性仍然会存在。如果抽样是无放回地进行的,样本必须要包含 n 个不同的元素,但是对于有替换的抽样,样本量为 n 的抽样可能包含小于 n 个不同元素。当抽样过程是被有放回地进行的,抽样方法被称为无限制随机抽样

(unrestricted random sampling)或者有放回的简单随机抽样(simple random sampling with replacement)。当抽样是无放回地进行的,这种方法则是无放回的随机抽样(simple random sampling without replacement)或者简称为简单随机抽样。使用随机数表进行的简单随机抽样需要忽略已经在样本中的重复抽中元素。与有放回的抽样相比,无放回的抽样能够给出更加精确的估计量(estimator),因此我们主要关注无放回的抽样方法。

在使用简单随机抽样选取了 250 名学生之后,假设我们现在已经收集完数据,并且我们对所有抽中的学生都给予了回应(无应答的问题我们会在第 9 章中介绍)。下一步,我们会通过总结个人的回应来对总体的某些特征进行推断,比如,每日平均看电视的时间和正在阅读小说的学生比例。此时,我们会再次介绍一些概念。依据抽样调查文献中的规范,我们用大写字母表示总体值和参数,用小写字母表示样本值和估计量。因此,Y_1, Y_2, …, Y_N 表示的是变量 y(比如,看电视的小时数)对总体中 N 个元素的取值,而 y_1, y_2, …, y_n 表示的是样本中 n 个元素的取值。一般而言,变量 y 对于总体中第 i 个元素的值是 Y_i($i=1, 2, \cdots, N$),而其对于样本中第 i 个元素的取值则是 y_i($i=1, 2, \cdots, n$)。总体的均值由如下公式给出:

$$\bar{Y}=\sum_{i=1}^{N} Y_i/N$$

样本的均值则是:

$$\bar{y}=\sum_{i=1}^{n} y_i/n$$

在抽样调查中，y 变量总体的方差一般被定义为：

$$S^2 = \sum_{i=1}^{N} (Y_i - \bar{Y})^2/(N-1)$$

而样本的方差则是：

$$s^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1)$$

然而，有些时候，总体方差被定义为其分母为 N 而不是 $N-1$，具体如下：

$$\sigma^2 = \sum_{i=1}^{N} (Y_i - \bar{Y})^2/N$$

（比如，$\sigma^2 = (N-1)S^2/N$）。

假设我们希望使用调查搜集到的数据来估计该学校所有学生平均每天看电视的时间 $\bar{Y}$。这时，我们就会考虑 $\bar{y}$ 是不是 $\bar{Y}$ 的一个足够好的估计量。由于 $\bar{Y}$ 是未知的，这一问题对于从一个特定样本得出的 $\bar{y}$ 而言也很难回答。然而，我们可以通过重复抽样得到的均值的性质来获得一些更可靠的估计。我们注意到，估计（estimate）指的是一个特定的值，然而估计量则是为了得到估计而使用的程序或者规则。在上面的例子中，我们可以通过均值的估计量 $\bar{y} = \sum y_i/n$ 来计算得到一个看电视的平均时间估计值 2.2 小时。统计理论提供的是评价估计量的方法，而非估计的方法。下面的论述会简要回顾统计推断理论中关于简单随机抽样的部分；对于统计推断的一个更加充分的讨论，请读者参考例如布莱洛克（Blalock，1972）等其他研究者的统计学著作。

在理论上，样本估计量的统计性质是建立在重复无限次

抽样过程的基础之上的。在目前的例子中，假设我们重复无限次从 1 872 名学生中抽取 250 人的简单随机抽样过程，并每次得到其样本均值（当然，每次抽取新的样本之前需要将前一次抽取的样本放回总体）。我们得到的样本均值的集合则会有一个分布，或称均值的抽样分布（sampling distribution）。如果样本量并不是过于小了——一般而言 10 或 20 就够了——统计理论证明这一分布近似于正态分布（normal distribution），并且这一分布的均值为总体均值 $\bar{Y}$。如果无限次重复抽取得到的样本估计的均值与总体均值这一参数相等，那么这一估计量则是总体参数的无偏（unbiased）估计量。因此，在简单随机抽样的情形下，$\bar{y}$ 是 $\bar{Y}$ 的无偏估计量。

尽管 $\bar{y}$ 的抽样分布以 $\bar{Y}$ 为中心，其任何一个具体的值可能会与 $\bar{Y}$ 不同。因此，我们需要一个描述不同估计值与 $\bar{Y}$ 接近程度的度量。一个常用的描述这一变异（variability）程度的度量是标准偏差（standard deviation），定义为方差的平方根。在这种情形下，标准偏差是样本均值分布的标准偏差。为了避免它与每一元素的值的标准误差相混淆，抽样分布的标准偏差一般被称为标准误差（standard errors）。我们将一个由 SRS 得到的样本均值标记为 $\bar{y}_0$（其下标 0 表示简单随机抽样），其标准误为 $SE(\bar{y}_0)$，以及标准误的平方，或者 $\bar{y}_0$ 的方差为 $V(\bar{y}_0)$。为了方便，大多数抽样误差的公式会写成方差的函数。一个样本量为 n 的 SRS 的样本均值的方差如下：

$$V(\bar{y}_0) = \frac{N-n}{N-1}\frac{\sigma^2}{n} \qquad [2.1]$$

或者等价的：

$$V(\bar{y}_0)=\left(\frac{N-n}{N}\right)\frac{S^2}{n}=(1-f)\frac{S^2}{n} \quad [2.2]$$

其中 $f=n/N$ 为抽样的比率。

这些公式表明 $V(\bar{y}_0)$依赖于三个因素：第一，$(N-n)/(N-1)$ 或者$(1-f)$，就是有限总体修正(Finite Population Correction, FPC)——当 N 非常大的时候，这两项的差距就非常小了；第二，样本量 n；第三，S^2 或者 σ^2，也即总体中变量 y 的方差。其中，FPC 表明调查的总体规模是有限的；而标准统计理论假设总体规模是无限的，从而人们可以进行无放回重复抽样。当总体的大小无限，或者进行有放回抽样时，就不会存在 FPC 这一项，从而公式 2.1 就可以简写为 $V(\bar{y}_0)=\sigma^2/n$ 了。因此，FPC 一项表示了无放回抽样相比于有放回抽样的益处。对于一个样本量大于 2 的样本，FPC 是小于 1 的，说明从简单随机抽样得到的 $\bar{y}$ 比从一个具有相同样本量的无限制样本中得到的 $\bar{y}$ 更加精确，或者有更小的方差。在很多实际情况下，如果总体非常庞大，即使样本很大，抽样的比率也还是很小的。在这种情况下，有放回和无放回抽样的区别就不再重要了，因为即使人们有放回的抽样，抽取某一特定元素多于两次的概率也很小。这一点能够用 FPC 表示出来。如果抽样比率(f)为 1/10，FPC 为 0.9，那么其对标准误差的影响为$\sqrt{1-f}=0.95$；如果 $f=1/20$，$1-f=0.95$ 从而$\sqrt{1-f}=0.97$。这些结果表明如果抽样比率是很小的，FPC 会非常接近于 1，因此对于标准误差的影响很小。当抽样比率比 1/20 甚至 1/10 小的时候，FPC 项一般都可以被忽略(当做 1)。

第二个影响 $V(\bar{y}_0)$的因素是样本量 n。这一点非常直

观，因为样本量越大，$V(\bar{y}_0)$越小。而相对不那么直观的一点是，对于庞大的总体，样本量在决定抽样的准确度上比抽样比率更能起到决定性的作用。举个例子，从一个具有 20 亿人的国家中抽取 2 000 人得到的结果，与从一个 4 万人的小城市中抽取 2 000 人的结果一样精确（假设两个总体的方差是相等的）。因此，对于越庞大的总体，抽样的益处就越明显。诚然，对于非常小的总体，抽样的好处可能并不大，即使 FPC 此时会发挥重要的作用。比如，在一个仅有 200 名学生的学校中调查全部人可能比抽取其中的 175 人更加方便。

第三个影响 $V(\bar{y}_0)$的因素是总体中 y 的方差，S^2 或者 σ^2。显然地，如果所有的学生看电视的量是相近的，那么任何样本的均值都会接近于总体均值。然而，如果他们在看电视的习惯上有很大差别，那么任意一个样本均值都可能会与总体均值的差异非常大。注意，S^2 或者 σ^2 都是总体参数；因此，在实际中我们并不知道它们的真实值。为了估计$V(\bar{y}_0)$，我们需要准确估计总体方差。使用 S^2 表示的公式 2.2 的好处在于，我们熟悉的样本估计量 $s^2=\sum_{i=1}^{n}(y_i-\bar{y})^2/(n-1)$ 是 S^2 的无偏估计量（但并不是 σ^2 的无偏估计量）。因此，$V(\bar{y}_0)$和 $SE(\bar{y}_0)$可以简单通过下面的公式来估计：

$$v(\bar{y}_0)=(1-f)\,s^2/n \qquad [2.3]$$

以及

$$se(\bar{y}_0)=\sqrt{(1-f)s^2/n} \qquad [2.4]$$

其中小写的 v 和 se 表示样本的估计量。

在已经估计了标准误差之后，我们可以计算总体均值的

置信区间(confidence interval)。比如,对于一个大样本,其对于 $\bar{Y}$ 的 95%置信区间是 $\bar{y}_0 \pm 1.96\text{se}(\bar{y}_0)$,其中 1.96 是从标准正态分布的表中得到的(95%的标准正态分布都落入围绕分布均值的 1.96 个标准偏差中)。举个例子,假设 250 个学生每天平均看电视的时间为 $\bar{y}_0 = 2.192$ 小时,其方差为 $s^2 = 1.008$,那么 $\bar{Y}$ 的 95%置信区间为:

$$2.192 \pm 1.96\sqrt{\left(1 - \frac{250}{1\,872}\right)\frac{1.008}{250}} = 2.192 \pm 0.116$$

就是说,我们对于区间 2.076 到 2.308 包含了总体均值有 95%的把握。

除了均值,人们关心的另一个参数是总体中有某一特征的人的比率(或百分比),比如目前阅读小说的学生的比例。比率的结果可以简单从均值得到,因为比率就是均值的一种特殊形式。为了说明这一点,我们假设当第 i 个元素有这一特征的时候 $Y_i = 1$,否则 $Y_i = 0$。因此,$\bar{Y} = \sum Y_i / N$ 就是总体中具有这一特征的人的比例 P,从而样本均值 $\bar{y}$ 就是样本比例 p。因此,一般而言,样本均值的性质也适用于比率。在我们目前讨论的 SRS 例子中,由于 $\bar{y}_0$ 是 $\bar{Y}$ 的无偏估计,因此 p_0 也是 P 的无偏估计。然而,由于变量 y 的取值仅仅是 0 和 1,其 S^2 和 σ^2 的表达式可以被简化为 $NPQ(N-1)$ 以及 $np_0q_0/(n-1)$,其中 $Q = 1 - P$,$q_0 = 1 - p_0$。使用这些表达式,我们有:

$$V(p_0) = (1 - f)\frac{NPQ}{(N-1)n} \qquad [2.5]$$

以及

$$v(p_0) = (1-f)\frac{p_0 q_0}{(n-1)} \qquad [2.6]$$

如果可以忽略 FPC,或者 n 非常大,$v(p_0)$就可以退化到 $p_0 q_0/n$ 的形式。这些公式也适用于百分比,只需修改为 $Q = 100 - P$ 以及 $q_0 = 100 - p_0$。

举个例子,假设样本中 250 个学生中的 165 人阅读小说,比如 $p_0 = 66.0\%$。那么 P 的 95%置信区间就是:

$$66.0 \pm 1.96\sqrt{\left(1-\frac{250}{1\,872}\right)\frac{66\times 34}{249}} = 66.0 \pm 5.5\%$$

即,我们有 95%的把握认为区间 60.5%到 71.5%包含了总体的学生读书的百分比。

先前的讨论回顾了根据 SRS 方法估计总体均值或者比率的步骤,以及计算相应的置信区间的方法。这对于从大样本来进行统计推断而言是标准的一般方法,其中唯一的区别在于 FPC 一项上。这一方法也可以被用到估计其他总体参数的过程中。

第3章

系统抽样

在前面的一章中，我们提到了使用随机数表来抽取一个包含 250 个学生的随机样本。这虽然是可行的，然而实现的过程却不免繁冗。另外，如果总体非常庞大，或者样本量增加，抑或所有的学生并非以身份码来识别的话，抽样的过程会更加耗费时间。系统抽样作为一种能够在很大程度上简化抽样过程的方法，近年来被广泛应用。系统抽样非常容易，因为它仅仅抽取一个随机起点后的每第 k 个元素。

一个很简单的例子是，假设我们从 2 000 个学生的学校中抽取一个有 250 人的样本。抽样比率为 250/2 000，或者 1/8。通过在 1 到 8 中随机选取一个数字，我们可以决定样本中的第一个元素，之后我们每隔 8 个学生抽取一人。如果随机数是 5，那么我们抽取的学生则是名单上的第五个、第十三个、第二十一个学生，以此类推。

在应用上一章节的例子时，系统抽样的过程比上文描述的要复杂一些，因为此时的抽样比率是 250/1 872 或者 1/7.488。因此，此时抽样间距(sampling interval)7.488 并非一个整数。有些时候，这个问题可以通过将这一数字四舍五入来处理，但相应的样本量也会有所变化。在这一例子中，7 人中抽 1 人将会得到一个样本容量为 267 或 268 的样本，

然而 8 人中抽一人则会得到容量为 234 的样本。如果人们不能够接受四舍五入的处理方法带来的样本量的变异性，我们还可以使用其他方法。一个方法是只保留间距的整数位（如 7 人中抽 1 人），然后从 1 872 名学生中随意选取一人开始，直到抽取到了我们希望的样本量（比如 250 人）。使用这一方法时，学生名单实际上被当做是环形的，所以名单中的最后一人之后接着是名单的第一个人。第二个方案是使用比率间隔（fractional interval），舍去小数位来决定抽样起点。在前面的例子中，人们抽取一个从 1 000 到 7 488 之间的四位随机数，比如 3654，作为抽样的起点。接下来，将这一数字的小数点前移三位，得到 3.654，然后从第三名学生开始抽取。将3.654重复加上 7.488，我们会得到 11.142、18.630、26.118 等。这一序列就会产生第十一、第十八、第二十六名学生，以此类推。因此，被选择的学生之间的间隔有时是 7，有时是 8。

一个识别系统抽样样本中的学生的方法是找出名单中的第三人、第十一人、第十八人，等等。另一个方法就是利用学生身份码。使用这种方法，人们可以将抽样间距累加，直到其超过 1917，也即身份码的最大值。在这一方法中，我们依然先舍去小数位来决定抽样的开始；但如果被选择的号码并不对应一位学生，这一号码就不会被选取。期望的样本量仍然是 250，但是实现的样本量则可能与 250 不同，因为其中可能出现号码与学生不对应的状况。

正如简单随机抽样一样，系统抽样给予总体中每一元素相同的抽取概率。比如，它是一个等概率抽样的设计。然而，与简单随机抽样不同的是，系统抽样中元素的不同集合

被抽取的概率并不相等。比如，在前面从 8 人中抽 1 人的例子中，元素 1 和 2 同时被抽取的概率为 0，而元素 1 和 9 同时被抽取的概率是 1/8，因为如果元素 1 被抽取，那么元素 9 必然被抽取。系统抽样的等概率抽样性质表明，样本均值是对总体均值的合理估计量。然而，由于系统抽样中不同元素的集合被抽取的概率并不相等，SRS 的标准误差公式并不能被直接应用到系统抽样中。

在从 8 人中抽 1 人的例子中，很容易确定均值或者比率的抽样分布：因为只有 8 个不同的样本，每一个样本都有相同的可能性出现，抽样分布就是八个样本的均值或者比率，每一个发生的概率为 1/8。系统抽样的一个局限在于，除非我们对名单的顺序作出假设，被抽取的元素值的变异程度并不能用来估计抽样分布的变异性的基础。为了说明这一点，我们再次考虑所有学生中阅读小说的学生所占比例这一问题。假设 2 000 名学生中有 1 500 个学生在阅读小说，而学校的学生名单的排列如果正好是按照每六个阅读小说的人之后是两位不阅读的学生的话，那么如果我们从 1 到 6 中随机选择起始点的话，抽取的样本就是包括全部阅读小说的学生（p＝100％）；然而如果我们从 7 和 8 中随机选择起始点时，则不会抽取到任何一个阅读小说的学生（p＝0）。因此，样本的估计量是非常不精确的，其真实的标准差为 43.3％。而每一样本的内部的变异性则为 0，因此对标准误差的大小没有任何指示作用。

为了对系统抽样下的估计量的标准误差作出估计，我们需要对总体做一些假设。有时候，我们可以假设名单在我们关心的变量上是接近随机的，那么随后抽到的样本就可以被

当做简单随机样本。按照字母顺序排列的名单一般可以被这样处理。有时候,名单可能是按照分组排序的(比如,不同年级的学生),而在不同年级中,我们关心的变量的变异程度可能有所不同。这时,系统抽样得到的样本就可以被当做分层样本(stratified sample,参见第 4 章)。抽样调查者常常会对名单中的顺序进行调整,然后进行抽取,从而得到按比率的分层抽样(proportionate stratification)结果。

正如上面所说的,当名单在我们关心的变量上具有一定的周期性,而抽样间距是这一周期长度的倍数时,系统抽样的效果会很差。但是,如果抽样的间距并不是这一周期长度的倍数,系统抽样的效果则尚可。为说明这一点,读者可以考虑上面学生阅读比率的例子中从 7 人中抽 1 人的情形。虽然抽样者需要警觉在周期性名单中系统抽样的潜在危险,这样的名单在实际中是很难碰到的,而一旦出现这种情形,也很容易被识别出来。在不过多担心名单序列的周期性的情况下,系统抽样可以被广泛应用。

第4章

分层抽样

一个经常会碰到的抽样调查的特征是总体中的一些元素信息是已知的。比如，在选择美国一个地区性样本的时候，该地区的地理位置的信息是已知的，比如它是否为内陆城市、郊区或者农村地区，同时人口普查的数据也可以提供关于这一地区的其他有价值信息——比如，之前人口普查中该地区的人口、人口变动率、在制造业中就业的人口比率，以及总人口中非白人的比率。抽样调查的设计阶段考虑到这些补充信息不仅能够提高样本设计质量，也可以在分析阶段来提高样本估计量的质量，或者两者兼有。这一部分会讨论利用补充信息，通过分层抽样技术来提高样本设计的质量。

分层抽样本质上是根据额外信息来将总体分为子总体(subpopulation)或者层(strata)，然后在每一层中分别进行抽样。分层抽样的好处在于，每一层中抽取的样本量是由抽样者控制的，而非由抽样过程随意决定的。通常，分层抽样得到的样本是从对应层的总体中按比例抽取的；换句话说，人们使用统一的抽样比率(sampling fraction)。因此，这通常被称为按比例分层(proportionate stratification)。然而，将总体划分为不同的层时并不需要按比例进行，非比例分层(dis-

proportionate stratification)同样也是可能的。在这一部分中，我们仅仅考虑层内使用的简单随机抽样，但是在之后我们会说明，其他抽样方法也可以被使用。

为了对分层进行处理，我们需要拓展前面引入的概念。具体而言，我们加入一个下角标 h 来表示对应的层 h。因此，N_h 就是层 h 的总体大小，而 n_h 就是在 h 层取的样本大小，$N=\sum N_h$ 以及 $n=\sum n_h$ 则是全部总体以及样本的大小；$f_h=n_h/N_h$ 是在层 h 中的抽样比率；$\bar{Y}_h$ 和 $\bar{y}_h$ 则是层 h 中的总体均值和样本均值；S_h^2 和 s_h^2 则是层 h 中总体元素的方差以及样本元素的方差。这里，我们引入一个新的符号 $W_h=N_h/N$，即层 h 在总体中所占的比重，并且 $\sum W_h=1$。

给定在每一层内进行的是简单随机抽样后，先前的结果就可以被应用到每一层中，那么对于每一个 h，$\bar{y}_h$ 是 $\bar{Y}_h$ 的无偏估计，它们的方差和标准误差可以根据公式 2.3 和公式 2.4 估计得到。分层抽样引入的新的问题在于，如何将不同层的均值组合起来得到对全部总体均值 $\bar{Y}$ 的估计量及如何来估计这个估计量的方差。对于前一个问题，我们可以将 $\bar{Y}$ 表示为 $\sum N_h\bar{Y}_h/N=\sum W_h\bar{Y}_h$。因此，对一个明显的 $\bar{Y}$ 估计量就是用层样本均值 $\bar{y}_h$ 代替未知的 $\bar{Y}_h$。因此，我们可以得到一个无偏估计量是 $\bar{y}_{st}=\sum W_h\bar{y}_h$，其中 st 表示分层。

如果我们在每一层中分别独立进行抽样，$\bar{y}_{st}=\sum W_h\bar{y}_h$ 的方差就可以根据公式 4.1 给出：

$$V(\bar{y}_{st})=\sum W_h^2V(\bar{y}_h) \qquad [4.1]$$

在每一层中使用 SRS，公式 4.1 可以表示为：

$$V(\bar{y}_{st}) = \sum W_h^2 (1 - f_h) S_h^2 / n_h \qquad [4.2]$$

上式由将公式 2.2 替换 $V(\bar{y}_h)$得到。对于 $V(\bar{y}_{st})$的一个估计量就是将 s_h^2 替换公式 4.2 中的 S_h^2：

$$v(\bar{y}_{st}) = \sum W_h^2 (1 - f_h) s_h^2 / n_h \qquad [4.3]$$

第1节 按比例分层

上面的公式对于不同层的样本都是适用的。在按比例分层的情况下,比如人们使用一个统一的抽样比率 $f_h = f$ 或者 $n_h/N_h = n/N$,这些公式可以被简化。按比例分层是一个等概率抽样设计,其中 $\bar{y}_{st}$ 可以被简化为简单样本均值:

$$\sum_h \sum_i y_{hi}/n$$

其中 y_{hi} 是层 h 中的第 i 个元素的 y 的值,这个总和是所有样本的元素相加得到的。此时,公式 4.2 中 $\bar{y}_{st}$ 的方差可以被简化为:

$$V(\bar{y}_{st}) = (1-f)\sum W_h S_h^2/n = (1-f)S_w^2/n \qquad [4.4]$$

其中 $S_w^2 = \sum W_h^2 S_h^2$ 是每一层方差的加权平均值。从而 $V(\bar{y}_{st})$ 可以由公式 4.5 估计:

$$V(\bar{y}_{st}) = (1-f)\sum W_h S_h^2/n \qquad [4.5]$$

在此,人们可能会发现按比例分层样本均值的方差(公式 4.4)与简单随机抽样的样本均值的方差公式(公式 2.2)非常相似。仅有的差别在于,简单随机抽样中的总体元素的方差 S^2 被按比例分层抽样中的不同层的方差的加权平均值 S_w^2 替代。对于具有很大样本量 N_h 的情形,我们可以近似将方

差分解为：

$$S^2 = S_w^2 + \sum W_h(\bar{Y}_h - \bar{Y})^2$$

由于上式的最后一项是非负的（平方项之和），从而S^2大于等于S_w^2。换句话说，按比例分层抽样的样本的精确度与具有相同样本容量的简单随机抽样差不多。给定总体的变异由于进行了按比例分层抽样，分层样本的均值的异质性越强，或者说在每一层内的元素的值更具一致性，就能得到比简单随机抽样更高的准确度。

正如之前我们讨论过的，简单随机抽样可以作为比较其他抽样方法的一个基准。一个经常被使用的比较的度量是设计效应（design effect），即复杂设计的抽样方法得到的估计量的方差与具有相同样本容量的简单随机抽样估计量的方差的比值。我们将估计量 z 的设计效应表示为：$D^2(z) = V(z)/V(z_0)$，因此按比例分层抽样的样本均值的设计效应为$D^2(\bar{y}) = S_w^2/S^2$，这一值在上面的近似下不超过 1。有时标准误差的比，比方差的比更合适；设计效应的平方根可以表示为$D(z)$。对 z 的设计效应的样本估计用 $d^2(z)$表示。另一个定义设计效应的方法是比较非限制样本（unrestricted sampling）而非简单随机样本。这一方法的优点在于可以比较复杂抽样样本的方差与标准情况下的方差。然而，由于简单随机抽样的方差与非限制抽样的方差的差别仅在于 FPC 项$(1-f)$，而其一般可以被忽略，这两个定义设计效应的区别是很微小的。

为了对按比例分层抽样进行说明，我们现在回到前面一章提到的例子。现在，我们假设学生名单被分为四个单独的名单，每一个对应一个年级（九年级、十年级、十一年级和十

二年级)。然后,我们按照年级进行分层抽取。表4.1的第二栏和第三栏给出了从每一年级的所有学生中抽取的人数和比重。第四栏给出了每一层中抽取的样本,其中统一的抽样比率为250/1 872,或者1/7.448。对于每天看电视的小时数,第五栏、第六栏和第七栏给出了其样本加总、样本均值以及每一层中样本的方差。第八栏和第九栏给出了每一层抽取到的阅读小说的学生数和比例。

表4.1 高中学生的按比例分层抽样(虚拟数据)

(1) 层	(2) N_h	(3) W_h	(4) n_h	(5) $\sum_i y_{hi}$	(6) $\bar{y}_h$	(7) S_h^2	(8) r_h	(9) p_h
九年级	524	0.28	70	168	2.40	0.941	35	50%
十年级	487	0.26	65	169	2.60	1.088	39	60%
十一年级	449	0.24	60	123	2.05	0.804	45	75%
十二年级	412	0.22	55	88	1.60	0.643	44	80%
总　计	1 872	1.00	250	548			163	

对于每天看电视的小时数,其样本均值可以通过$\sum W_h \bar{y}_h$计算。然而由于样本是按比例得到的,这一值也可以通过简单的样本均值算出来:

$$\bar{y}_{st}=\sum_h \sum_i y_{hi}/n=548/250=2.192$$

以相同的方式,样本中阅读小说的学生的比例可以用$p_{st}=100\sum r_h/n=100(163/250)=65.2\%$计算。$\bar{y}_{st}$的方差可以根据公式4.5计算如下:

$$v(\bar{y}_{st})=\left(1-\frac{250}{1872}\right)\left(\frac{0.8808}{250}\right)=0.003053$$

并且$se(\bar{y}_{st})=0.0553$。因此,对于$\bar{Y}$的95%置信区间是$\bar{y}_{st}\pm 1.96se(\bar{y}_{st})$,或者2.08到2.30。

p_{st}的估计方差可以从公式 4.5 中得到，注意，$S_h^2 = n_h p_h q_h/(n_h - 1)$ 表示比例。从而得到：

$$V(p_{st}) = \left(1 - \frac{250}{1\,872}\right)\left(\frac{2\,160}{250}\right) = 7.486$$

并且 $se(p_{st}) = 2.736\%$。

对 P 的一个 95%置信区间因此可以是 $p_{st} \pm 1.96se(p_{st})$，或者为 59.8%到 70.6%。

$\bar{y}_{st}$和 p_{st}的设计效应可以由S_w^2/S^2 估计。对于 $\bar{y}_{st}$，S^2 近似为 1.008(这里没有给出计算过程；请参考 Cochran，1977：第 5A11 节)，从而 $d^2(\bar{y}_{st}) = \frac{0.8088}{1.008} = 0.87$。换句话说，我们需要一个容量为 250/0.87 =286 的简单随机抽样的样本才能达到相同的精确度。这一在精确度上的增加的来源是不同年级看电视时间明显的变异性。

对于 p_{st}而言，s^2 近似为 2 278，从而 $d^2(p_{st}) = 0.95$。此时，使用分层抽样得到的精确度的增加比我们期望的要小，特别是当我们考虑到表 4.1 第九列中四个层的比重的明显差异时。然而，这一相对小的精确度增加是一个惯例，除非一些层具有非常高(比如高于 90%)或者非常低(低于 10%)的比重。

第2节｜非比例分层

由于按比例分层能够提供比较简单的估计量，同时保证其精度不低于简单随机抽样，这一方法被经常使用。然而，有时非比例分层对我们也是有帮助的。

非比例分层的一个重要目的是，在给定的资源下实现样本估计量的精确度的最优化。为了达到这一目的的最优分配是使分层抽样中的抽样比率与该层中元素的标准偏差成正比，并且与每加入来自该层的一个元素所需要的成本的平方根成反比，比如 $f_h \propto S_h/\sqrt{c_h}$，其中 c_h 是层 h 中每一样本元素的成本。正如人们可以预计的，这一结果表明更加具有异质性的层以及成本比较低的层应当以更高的比率被抽取。抽样成本常常并不随层而变化，从而最优的资源配置可以写为 $f_h \propto S_h$，或被称做内曼配置(Neyman allocatio)。

实现最优配置的一个实际性难题在于人们缺乏对层中元素的方差和抽样成本的深入了解。但幸运的是，一般而言，一个合理的、相对精确的估计就够了，因为略微偏离最后配置导致的准确度的缺失不是很大。另一个困难在于抽样调查的多重目的性，因为对于某个变量的最优配置往往对另一个变量而言是相当糟糕的。与按比例分层不同，非比例分层的估计量的精确度可能会低于具有相同样本的简单随机

抽样的估计精度。

非比例抽样的另一个用途在于对某一层分配足够的样本量，从而对这一层的估计的精确程度能够得以保证。通常人们不仅需要对总体的样本估计值，也需要一些子总体的样本估计值，或称做研究领域(domains of study)。当一个样本量较小的层代表一个研究领域时，比例分配可能会使得本层的样本量太小，从而不能产生具有足够精度的估计；对此的一个弥补方式是，在这一层中采取更高的抽样比率。

然而，人们需要使用非比例抽样的另一个情形就是，该抽样调查的目的是为了在不同层的估计值之间进行比较，而不是将它们合并起来进行一个总体的估计。比如，高中调查的目的可能是比较不同年级看电视的小时数，而不是估计所有年级看电视的小时数。当我们仅有两个层的时候，为了估计两个层均值差别的估计量的最优配置如下：

$$\frac{n_1}{n_2}=\frac{S_1/\sqrt{c_1}}{S_2/\sqrt{c_2}}$$

如果层的方差与成本是相等的，最优配置可以近似简化为 $n_1=n_2$。需要注意的是，为了在不同层之间进行比较，虽然层的总体规模是不相关的，但它们在对全部总体的性质进行估计的时候很重要。当我们同时需要在层之间比较以及实现对全部总体估计的时候，层的规模上的较大差异可能导致样本分配的冲突。比如，如果层的方差以及成本都是相等的，第一层包含了 90%的总体而第二层仅有 10%，那么使用 500 个元素的样本估计全部总体的均值的最优配置就是在第一层中抽取 450 人，在第二层中抽取 50 人。但是，如果我们

的最优配置是为了估计两个层均值之间的差别，那么应该在两个层之中分别抽取250人。当这种情形出现的时候，对于一个目标的最优配置可能对另一个目标而言有很大影响，但有些时候，人们可以使用一些折中的方法。

为了给出一个非比例抽样的例子，我们重新回到前面的高中案例。表4.2的数据与表4.1对应，我们重新安排了第四栏的内容以使其符合非比例抽样。我们选择的分配是将250的样本尽量平均地在四个层之间划分，因为每一层可能被独立地选为研究的对象(假设层中的元素的方差与成本都是相等的)。

表4.2 高中学生的非比例分层抽样(虚拟数据)

(1) 层	(2) N_h	(3) W_h	(4) n_h	(5) $\sum_i y_{hi}$	(6) $\bar{y}_h$	(7) S_h^2
九年级	524	0.28	63	151.2	2.40	0.941
十年级	487	0.26	63	163.8	2.60	1.088
十一年级	449	0.24	62	127.1	2.05	0.804
十二年级	412	0.22	62	99.2	1.60	0.643
总　计	1 872	1.00	250			

整体而言，看电视时间的均值可以按照公式 $\bar{y}_{st}=\sum W_h\bar{y}_h$ 计算出来，为2.192，与前面的值相等。在这个例子中，简单均值 $\sum_h\sum_i y_{hi}/n=2.165$ 并不是 $\bar{Y}$ 的有效估计量：在本样本中被过度代表的年级越高，得出的报告显示看电视的时间也越少，简单均值从而低估了 $\bar{Y}$。通过将总体中层的比例 W_h 作为权重，加权均值能够修正样本的不均衡性。$\bar{y}_{st}$的方差，根据公式4.3的估计，为 $v(\bar{y}_{st})=0.003117$，从而 $se(\bar{y}_{st})=0.0558$。比较这一标准误与从按比例抽样得到的标准误(0.0553)，我们发现非比例抽样产生的对总体均值的估计量的精度更小。

第 3 节 | 层的选择

标准分层需要满足两个条件:第一,每层占总体的比例,也即 W_h,需要是已知的;第二,从每一层中分别抽取样本是可能的。如果这些条件得不到满足,则需要新的技术来处理,我们会在后面进行讨论(参见第 10 章关于事后分层以及第 7 章中两阶段抽样的内容)。但如果这些条件能够满足,我们处理不同的层时就有很大的灵活性。在这里仅需要注意的一个限制是,每一层中至少需要选取一个,否则试图得到计算总体均值的无偏估计量是不可能的。如果我们也希望从样本中得到对标准误的估计,每一层中必须至少选取两个。

在实践中,人们常常有关于总体的可观信息,从而可以将其用到分层的过程中,进而为分层提供更多的灵活性。具体如何操作是由分层的目的决定的。为了获得整体估计量的高精度,我们选取的层需要尽可能地具有较强的内部同质性(对于我们关心的变量而言)。如果我们关心不同层对应的小的领域中的分别估计,那么每一个领域需要对应一个层——或者一组层——人们可以将这些层用较高的比率进行抽样来达到目标的样本量。有时,在一个层中使用不同的抽样方法可能是有帮助的。比如,在调查一个小城市及其周

边农村地区的人口时,人们可以考虑在城市层中使用系统抽样而在农村层中进行地区抽样。

我们回到高中的例子来说明如何将不同的抽样方法结合起来。现在,我们假设除了年级以外,还会在分层中使用到以下变量:学生的性别、学业表现的总体成绩(被分为高、中、低三个层次)以及居住地点(被分为三类),每一个变量都被认为与学生的看电视量有关。对于居住地点而言,其原因是居住地对应了不同的住房,它能够作为一个家庭社会阶层的代理变量,从而与看电视量有关。在使用这些变量分层的时候,我们不需要一些客观规则。相反,不同层的设置可以以主观的方式进行,而这样做也不会使调查的估计量出现偏差;每一层内的按比例抽样可以防止选择偏差的出现。衡量层的设置是否成功的准则是其内部的同质性,而这显然影响到调查估计量的标准误。在这个例子中,性别可能在解释九年级和十年级看电视量的差别中并不重要,而在解释十一年级和十二年级的差别中非常重要。另外,几乎所有十一年级和十二年级的学生可能最终都具有同样的居住地,使得这一变量不是区分高年级学生的重要因素。考虑到这些问题,我们会这样分层:首先,将学生按照年级分层;然后,在十一年级和十二年级内部,将学生按照学业表现分为三级别,然后再细分为男生和女生;在九年级和十年级内部,将学生按学业表现分为三个级别,然后按照居住地的三类进行细分。这一过程会产生 30 个层。如果有一些层中的样本量太小了(比如少于 15 个学生,这个数字能够保证按比例抽样的 250 个元素的样本中能够有两个选择),我们可以将其与相邻的层合并。

然而，在按比例将样本分为众多层的过程中，可能会出现一个问题，就是一些层需要的样本量可能很小而且零碎。比如，将要求的抽样比率 1/7.448 应用到只有 19 个高中生的层中，会产生一个仅有 2.54 个学生的样本。虽然将大样本量近似为其相邻的整数对抽样的可能性的影响仅是微乎其微的，但这对于小的样本量并不成立。一个解决这一问题的一般性方法是使用模糊（implicit）而不是精确（explicit）抽样的方法。模糊分层包含了将总体的元素按照层列举，然后在名单中整体使用系统抽样方法。按照这一方法，在具有 19 个学生的层中，取决于步骤开始时选择的随机数，可以得到 2 个或者 3 个学生。

第5章

整群抽样和多阶抽样

在大多数抽样问题中，总体可以被当做由一些元素的组构成的集合。一种抽样方法就是将这些组当做层，正如我们在上一节中介绍的。在这种情形下，人们从每一组中分别抽取样本。另一种抽样方法则可以将这些组作为群，在抽样调查中抽取其中的一部分。如果所抽取的群中的所有元素都被包含在样本中，这种方法就被叫做整群抽样(cluster sampling)。如果从每一个选取的群中抽取一些元素作为样本，这种方法就叫做两阶段抽样(two-stage sampling)。人们会常常用到群的阶层(hierarchy)：首先选取一些大的群，然后在所选的大的群中抽取一些小的群，如此进行直到最后的元素是从最后一阶段的群中抽取出来的。因此举个例子来说，为了调查一个国家的学生，人们可以首先选取一些学校，然后从选取的学校中抽取班级的样本，最后从所选的班级中抽取学生样本。这种方法一般被叫做多阶抽样(multistage sampling)，尽管有时候也被叫做整群抽样。

尽管层和群都是元素的分组，但是它们所服务的抽样目标是完全不一样的。由于在样本里完全代表层，如果层中的同质性(对于所关注的变量而言)比较强的话，这一方法具有优势。然而对于抽取的群而言，抽取的部分必须能够代表没

有被抽取的部分;而当每一个群中的元素的异质性(对于所关注的变量而言)比较强的话,这一方法比较有优势。按比例分层抽样一般可以实现比较高的精度。然而,除非在一些特殊情况下,相比于具有相同样本量的简单随机抽样,整群抽样则会损失一些精度。人们使用整群抽样,是因为这一方法无论在抽样还是在搜集数据时都比较经济。但是,如果使用整群抽样得到的好处无法弥补精度上的损失,那么使用整群抽样就是不合适的。

在这一节中,为了简单起见,我们设定一个不那么现实的假设:所有的群都具有相同的规模大小,也即 B(下一节会讨论不等规模的群)。从整体中的所有 A 个群中,人们用简单随机抽样抽取到了 $\underline{a}$,并且所选的群中列举了所有元素(注意:只要我们在文中使用了带下划线的 a,就表示它被用做一个数学符号而不是文字)。样本量从而是 $n=aB$,抽样比率为 $f=n/N=aB/AB=a/A$。在总体中,令 $Y_{\alpha\beta}$ 表示群 α 中的元素 β,令

$$\bar{Y}_{\alpha}=\sum_{\beta}^{B}Y_{\alpha\beta}/B$$

表示群 α 的均值,并且令

$$\bar{Y}=\sum_{\alpha}^{A}\sum_{\beta}^{B}Y_{\alpha\beta}/N=\sum_{\alpha}^{A}\bar{Y}_{\alpha}/A$$

为总体均值。在样本中,对应的量表示为:

$$\bar{y}_{\alpha}=\sum_{\beta}^{B}y_{\alpha\beta}/B$$

以及

$$\bar{y}_c = \sum_{\alpha}^{a}\sum_{\beta}^{B} y_{\alpha\beta}/n = \sum_{\alpha}^{a}\bar{y}_\alpha/a$$

在群大小相等的情形下，总体均值是 A 个群的均值的简单平均，而样本均值是 a 个抽取的群的均值的简单平均。因此，群的简单随机抽样可以被当做从 A 个均值的总体中抽取 a 个均值。因此，我们立即可以得到 $\bar{y}_c$ 是 $\bar{Y}$ 的无偏估计量，并且其方差可以从公式 2.2 得到如下：

$$V(\bar{y}_c) = \left(1 - \frac{a}{A}\right)\frac{S_a^2}{a} \qquad [5.1]$$

其中

$$S_a^2 = \sum_{\alpha}^{A}(\bar{Y}_\alpha - \bar{Y})^2/(A-1)$$

为群均值的方差。另外我们也可以得到：

$$v(\bar{y}_c) = \left(1 - \frac{a}{A}\right)\frac{s_a^2}{a} \qquad [5.2]$$

是 $V(\bar{y}_c)$ 的一个无偏估计量，其中

$$s_a^2 = \sum_{\alpha}^{a}(\bar{y}_\alpha - \bar{y})^2/(a-1)$$

将 $V(\bar{y}_c)$ 和从一个样本量为 $n = aB$ 的简单随机抽样中的均值的方差进行比较，我们得到 $\bar{y}_c$ 的设计效应为：

$$D^2(\bar{y}_c) = \frac{S_a^2/a}{S^2/aB} = \frac{BS_a^2}{S^2}$$

$D(\bar{y}_c)$ 的大小取决于 S_a^2 与 S^2 的比值，而这一比值则取决于群的选取。比如，假设总体中群的个数 A 比较大，人们从中随机抽取群，那么 S_a^2 则是简单随机抽样中 B 个元素的均

值的方差,可以近似表示为S^2/B。在这些前提下,$D^2(\bar{y}_c)=1$。但如果人们选取的群具有更高的内部同质性,而非像随机选取的那样,那么群的均值就会相应具有更强的异质性;从而S_a^2将比S^2/B大,因此$D^2(\bar{y}_c)$大于1。

另一个有用的表示群样本的均值的设计效应的方法是:

$$D^2(\bar{y}_c) \approx 1+(B-1)\rho \qquad [5.3]$$

其中ρ是层内的相关系数,衡量的是群内部的同质性程度(Kish, 1965:第5.4节)。在一个较大的总体中,如果群是随机形成的,那么ρ近似于0;因此$D^2(\bar{y}_c)=1$。ρ为负值时,说明相对随机形成的群而言,此时的群具有更强的内部异质性,但是ρ不会比$-1/(B-1)$更小。一个负的ρ会导致设计效应小于1,说明整群抽样比简单随机抽样更加精确。然而在实际中,负值的ρ很少出现。一般而言,ρ是比较小的正数(基本都小于0.15),从而$D^2(\bar{y}_c)>1$。ρ的最大可能取值为1,对应每一个群中所有元素都具有相同的取值的情形。

我们再次回到高中的例子。现在我们假设,在某一时刻学校包括$A=78$个班级,而每个班级有$B=24$名学生,只让选中的班级中的学生填写问卷是很轻松且经济的。一个包含10个班级的样本,会包含240名学生。下面的数字代表10个班中汇报阅读小说的学生的比例:

$$\frac{9}{24},\ \frac{11}{24},\ \frac{13}{24},\ \frac{15}{24},\ \frac{16}{24},\ \frac{17}{24},\ \frac{18}{24},\ \frac{20}{24},\ \frac{20}{24},\ \frac{21}{24}$$

整体的比例为$p_c=160/240=66.7\%$。从公式5.2中,我们令$p_\alpha=\bar{y}_\alpha$,且$p_c=\bar{y}_c$,然后得到:

$$v(p_c)=\left(1-\frac{10}{78}\right)\frac{0.02816}{10}=0.002455$$

从而 se(p_c)为 0.04955 或者 4.96%。因为 s_a^2 具有 9 个自由度,我们应当使用 t 分布而不是正态分布来构建总体中百分比的置信区间。因此,一个 P 的 95%的置信区间为 66.7 ± 2.26(4.96)或者 55.5%到 77.9%,其中数字 2.26 是自由度为 9 的 t 分布的 95%的分界点。

我们可以将 v(p_c)的值与具有相同样本量的简单随机抽样得到的阅读比例的方差进行比较。根据公式 2.6,我们得出:

$$v(p_0) = \left(1 - \frac{240}{1\,872}\right)\frac{0.6667 \times 0.3333}{239} = 0.0008106$$

从而 se(p_0) = 0.02847 或者 2.85%。整群抽样得到的样本比例的设计效应是:

$$d^2(p_c) = 0.002455/0.000810\,6 = 3.029$$

使用公式 5.3,我们可以得到对 ρ 的一个估计:

$$\hat{\rho} = [d^2(p_c) - 1]/(B - 1) = 0.088$$

这些结果表明,正的内部相关性导致整群样本比具有相同样本量的简单随机抽样样本具有更少的精度。近似地,忽略 FPC 项的影响,整群样本要是简单随机抽样样本量的三倍才能达到相同的精度。

正如公式 5.3 所说明的,整群样本的均值的设计效应取决于两个因素,内部相关系数 ρ 以及群的大小 B。上面例子中产生较大的设计效应的原因是在班级内部看电视的时间具有很大的同质性,或者群间也就是班级之间有较大的异质性。即使 ρ 比较小,在($B-1$)很大的情况下,设计效应也会很大。如果班级的规模为 8 而其中的内部同质性保持不变,设计效应就可以减少到 1.62。在实际中,随着群的规模的减

少，ρ 的值一般会增加，但变化的速度较慢，因此 B 的减少会主导设计效应的变化。

在整群抽样的情形下，这一论点就意味着，给定抽到的群足够大以便节省调查和数据搜集的成本，群的样本量越小越好。如果我们有一个群的阶层(hierarchy of clusters)，能够满足所需条件的规模最小的群是更好的选择。在高中的例子中，学生可以按年级或者班级分成小组；但这里年级作为群的单位就太大了，所以班级是更好的选择。整群抽样的问题在于，因为群一般包含了由于其他原因形成的小组，即使是在最低水平，所得的群也常常太大从而很难在整群抽样中得到有效的利用。这一问题的一个明显的解决方案是，将这些群分成子群(subclusters)用来抽样，这就是多阶抽样的基本方法。

考虑一个两阶段抽样，从 A 个群的总体中通过简单随机抽样抽取出 a 个群来，然后人们在每个抽到的群中用简单随机抽样从 B 个元素中抽取 b 个单位。因此简单样本均值为：

$$\bar{y}_{ts} = \sum_{\alpha}^{a} \sum_{\beta}^{b} y_{\alpha\beta}/n = \sum_{\alpha}^{a} \bar{y}_{\alpha}/a$$

仍然是总体均值的无偏估计，但是现在需要注意：

$$\bar{y}_{\alpha} = \sum_{\beta}^{b} y_{\alpha\beta}/b$$

是对于群 α 的样本均值，而不是该群的真实均值，因为它并不是一个彻头彻尾的整群抽样。$\bar{y}_{ts}$的方差为：

$$V(\bar{y}_{ts}) = \left(1-\frac{a}{A}\right)\frac{S_a^2}{a} + \left(1-\frac{b}{B}\right)\frac{S_b^2}{ab}$$

其中

$$S_b^2=\sum_{\alpha}^{A}\sum_{\beta}^{B}(Y_{\alpha\beta}-\bar{Y}_{\alpha})^2/A(B-1)$$

是群内部元素均值的方差。这一公式中的第一项代表整群抽样，而第二项表示在所选群内再次抽取子样本带来的新的方差。如果 $b=B$，那么第二项为 0，从而公式回到公式 5.1 给出的整群抽样的均值方差公式。如果 $a=A$，所有的群都包含在样本中，那么它们就成为了层。当 $a=A$ 时，第一项为 0，然后第二项就是公式 4.4 中按比例分层抽样的均值的方差表达式：$f=b/B$，$n=ab$，并且 $S_w^2=S_b^2$。

$V(\bar{y}_{ts})$的一个无偏估计量如下：

$$v(\bar{y}_{ts})=\left(1-\frac{a}{A}\right)\frac{s_a^2}{a}+\frac{a}{A}\left(1-\frac{b}{B}\right)\frac{s_b^2}{ab}$$

其中

$$s_a^2=\sum_{\alpha}^{a}(\bar{y}_{\alpha}-\bar{y})^2/(a-1)$$

并且

$$s_b^2=\sum_{\alpha}^{a}\sum_{\beta}^{b}(y_{\alpha\beta}-\bar{y}_{\alpha})/a(b-1)$$

这一公式看起来有点繁冗，因为 s_b^2 包含了在每一抽取的群中计算元素的方差。如果第一阶段的抽样比率 a/A 非常小，那么 $v(\bar{y}_{ts})$的第二项也非常小；所以在近似的时候，这一项可以被去掉。从而我们可以得到如下估计量：

$$v(\bar{y}_{ts})=s_a^2/a \qquad [5.4]$$

这样计算起来就简单多了。本质上，这一近似是将第一阶段的抽样当做有放回抽样。给定第一阶段的抽样比率比

较小,而事实上也的确如此,这一近似是可以接受的。这一近似在复杂抽样设计中被广泛应用,并且在很多计算抽样误差的电脑程序中也会用到。

另一个对两阶段抽样的近似则考虑到了在概念上总体中的每一个群都可以被划为 B/b 个末级群(Ultimate Clusters, UCs),每个末级群包含 b 个元素。使用这一设计,我们在每个群中使用简单随机抽样,末级群就可以被当做在每一群中由简单随机抽样得到的。在每一个群中,首先使用简单随机抽样抽取 b 个元素来组成第一个 UC,然后再使用简单随机抽样从剩余的元素中继续抽取 b 个元素,如此往复直到我们有 B/b 个UCs,而所有的元素都被包含进去(为简单起见,我们假设 B/b 是整数从而最后一个 UC 也包含 b 个元素)。然后我们从 AB/b 个末级群的总体中使用简单随机抽样抽取了末级群,其中选择的末级群中的所有元素都被包含在样本中。这一设计非常接近于上面讨论的两阶段抽样设计。与两阶段抽样不同的是,在第二阶段抽样中仅仅从每个选中的群中抽取一个末级群,而这一限制对于末级群抽样而言并不存在。然而,给定 a/A 非常小,从一个群中抽取两个 UCs 的几率非常小;在这种情况下,末级群抽样对于两阶段抽样是一个比较好的近似。末级群抽样的吸引力在于其简洁性:它是一个完整末级群的样本,并且适用于整群抽样的公式也同样适用于此。因此,一个近似的对 $\mathrm{V}(\bar{\mathrm{y}}_{\mathrm{ts}})$ 的估计就可以直接从公式 5.1 得到如下公式:

$$\mathrm{v}(\bar{\mathrm{y}}_{\mathrm{ts}})=\left(1-\frac{\mathrm{ab}}{AB}\right)\frac{\mathrm{s}_{\mathrm{a}}^{2}}{\mathrm{a}} \qquad [5.5]$$

其中抽样比率为 $\mathrm{a}/(AB/\mathrm{b})=\mathrm{ab}/AB=n/N$,并且 $\mathrm{s}_{\mathrm{a}}^{2}$ 为抽取的末级群样本均值的方差。在近似于两阶段设计的情

形下(比如,a/A 非常小),FPC 项可以被去掉,从而产生与有放回近似相同方差的估计量。

对于末级群抽样设计而言,样本均值的设计效应由公式 5.3 给出,也即 $1+(b-1)\rho$,其中 ρ 为 UC 内的相关系数而 b 为UC 的样本量。当我们使用简单随机抽样抽取 UC 时,它们的期望同质性与其原始的群的同质性是相同的。因此,作为一个近似,两阶段样本的设计效应为:

$$D^2(\bar{y}_{ts}) \approx 1+(b-1)\rho \qquad [5.6]$$

正如公式 5.6 说明的,如果 ρ 为正值,设计效应则随着子样本量 b 的减少而减少:对于一个固定的总样本量 $n=ab$,子样本量越小,被抽取的群的数量就越大,从而样本均值就更加精确。然而,群之间的样本越分散,抽样调查的成本也就越高,从而对于一个固定的预算而言,我们会得到更小的样本。人们往往需要平衡这两个因素,来决定抽取的群的数量 a 与每一个群中包含的元素 b 的最佳组合。为了实现这一点,我们需要设定一个调查成本的结构。一个简单的模型是 $C=aC_a+nc$,其中 C 为总成本,C_a 为每个群的成本,c 为每一个元素的成本。在这个模型下,最优的 b 可以近似地由最小化样本均值的方差得到(Kish, 1965:8.3B 部分)

$$b_{opt} \approx \sqrt{\frac{c_a}{c}\frac{(1-\rho)}{\rho}} \qquad [5.7]$$

从上面的公式我们知道,在其他条件保持不变的情况下,如果群内的同质性越强,每个元素的成本就越高,群的成本越低,那么样本就应当在群之间更加分散(比如,b 取比较小的值)。如果相对成本 C_a/c 约为 17,同时 $\rho=0.07$,那么

$b_{opt}=15$。给定总的预算，抽取的群的数量就可以被决定。

虽然上面的成本模型过于简单，但它基本能够给我们一个一般性的指导。人们当然可以使用更加复杂的模型，只是需要考虑更多的复杂性是否值得。即使对于简单的模型，成本构成的估计也是非常复杂的。除了成本以外，对 b_{opt} 的估计也需要对 ρ 的估计。这一估计常常基于过去的包含类似变量和抽样设计的调查。由于抽样调查具有多重目的，不同的变量可能会带来非常不同的 ρ 的取值，而对 b 的选取在一定程度上则是对不同研究目的的一种折中。

人们使用多阶抽样一般是为了其在抽样和数据搜集方面的经济性。使用地区抽样(area sampling)方法带来的经济性可以是很可观的，其中只有到最后阶段需要抽取群的时候才需要整合元素列表(比如，城市街区或者更小的单位)。通过面对面的访问来搜集数据，多阶抽样能够大大地减少访问员的行程。如果总体非常庞大并且非常分散，单阶段的样本可能太过分散，而多阶抽样则能够集中于一些地区进行访问。整群抽样则不能够明显提供使用电话访问或者邮件访问时的数据搜集方面的经济性的优势(除非人们使用面对面的访问来跟踪调查或者解决无应答问题)，但使用电话访问时整群抽样则可以具有抽样上的经济性(参见第12章)。

假设人们希望在某个城市中执行一个面对面访问的家庭调查。如果该城市比较小而且住户的名单是可以得到的，那么一个单阶段的按地区或其他变量的比例分层抽样可能是最佳的选择。但是如果城市比较小，而且我们没有住户的名单，那么人们可以使用两阶段的抽样设计来节省抽样成本；按照城市街区的分层样本可以在第一阶段中产生，然后

可以对住户进行标记并抽样。在大城市中,即使可以得到名单,两阶段的样本仍然是更好的选择,因为可以节省访问员的时间和成本。总体的规模越大,人们越可能使用更多阶段的抽样方法。对于一个对全体美国人的调查,人们一般需要三阶段或者更多阶段的抽样(参见第12章)。

为了简单起见,前面的讨论假设群和元素都是由简单随机抽样抽取的。在实践中,在有所需的分层信息时,多阶抽样使用的都是分层抽样,而系统抽样也常常被用到。分层抽样在抽取群的时候比抽取元素的时候更加重要,因为它在抽取群的时候能够带来更高的精度。另外,许多分层因素一般在对群分层时都适用。对多阶抽样中的第一阶段的群进行分层,或称初级抽样单位(Primary Sampling Units, PSU),使得人们得到尽可能多地进行PSU,然后从每一层(或者使用系统抽样的有序名单时的模糊层)中选取一个PSU。有时候,级抽样单位的样本是进一步由控制选择技术(technique of controlled selection)来控制的(Goodman and Kish, 1950; Hess et al., 1975)。

当我们从一个层中选取单个的级抽样单位,即PSU时,无法直接估计层内部的方差。为了让人们可以估计抽样误差,我们一般将一对相似的层合并起来,并将其当做每一对构成一个更大的层。这种折叠层法(collapsed strata)会导致对样本误差的高估,但是如果一对层是非常接近的话,那么高估就不是严重的问题。如果人们将一对层打散,在每个中间进行初次选取,然后可以在每个层中得到两个PSU,这种方法一般被称为配对选取(paired selection design)。被打散的层中的初级抽样单位,即PSU,一般被当做有放回的抽取,因此可以使用简单的有放回方差的估计量。

第6章

按规模大小成比例的概率抽样

上一部分中我们假设群的规模是相等的，然而这一假设在实践中很难被满足。事实上，由于实际中的群的规模几乎总是在变化，因此抽样者主要利用自然的分组来设定群。不可能所有的高中的班级都包含 24 个学生，其人数可能在 20 到 30 之间；街区也会有不同个数的住户，区县也是如此（这一单位经常被用到美国的全国调查 PSU 中）。我们下面会解释这一规模的变化给我们带来的问题，以及克服这些问题的一些方法。

为了便于说明，我们会举一个虚拟的例子。我们要从一个包含 9 个街区的总体中（初级抽样单位，即 PSU）选取住址的一个等概率抽样样本，可以将这一例子设想为是在更大的设计中分层。9 个街区包含了 315 个住址，其中我们希望抽取一个容量为 21 的样本，这意味着总体的抽样比率为 1/15。在抽样的第一阶段，我们需要抽取 3 个街区，然后从所选的街区中抽取住户。一开始，我们假设街区的规模，比如每个街区包含的住户数量是已知的且没有误差的。我们用 B_α 表示在街区 α 的住户的数量，具体如下：

街区：	1	2	3	4	5	6	7	8	9	总计
B_α：	20	100	50	15	18	43	20	36	13	315

可以考虑的第一种抽样方式是在第一阶段用简单随机抽样抽取三个街区。每个街区从而有 3/9 =1/3 的概率被抽到。使用概率的乘法定律我们可以得出选取到某一家庭住址的概率。一般而言,使用两阶段的设计,抽到街区 α 中住户 β 的概率如下:

$$\mathrm{P}(\alpha\beta)=\mathrm{P}(\alpha)\mathrm{P}(\beta \mid \alpha) \qquad [6.1]$$

其中 $\mathrm{P}(\alpha)$是街区 α 被抽到的概率,$\mathrm{P}(\beta|\alpha)$为第二阶段中抽到街区 α 中的元素 β 的概率,给定第一阶段抽到了街区 α。这一公式可以在必要的时候被拓展为多阶段抽样的公式;在抽样调查文献中,一般称其为选择方程(selection equation)。

目前的例子要求一个 $f=\mathrm{P}(\alpha\beta)=1/15$ 的等概率抽样设计。由于群被选取的概率相同,$\mathrm{P}(\alpha)=1/3$,从而我们可以得到 $\mathrm{P}(\beta \mid \alpha)=1/5$。换句话说,在每一群中的第二阶段的抽样比率为 1/5。现在,考虑一些根据这一抽样设计可能得到的样本。在一个极端,被抽到的街区可能包含了最少的元素——街区 4、5 和 9——然而在另一个极端,它们可能由最大的三个街区构成——2、3 和 6。在前者的情况下,在每个群内使用 1/5 的抽样比率总共会产生 9 个家庭住址,而在后者的情形下,则会包含 39 个家庭住址。平均而言,如果我们考虑到所有可能被抽到的三个街区的组合,样本量为 21,然而实际中根据一次抽取得到的样本量则可能与这一数字有很大差别。

这个例子中的样本量的变异性很大程度上源自一些抽取的群所包含的元素太少。然而,在其他抽样过程中,群的

规模的变化则可能比本例中街区群的规模变化大得多。因此，很明显我们需要一种方法来对抽取的样本量的潜在变化进行控制。虽然我们并不需要把样本量固定在某个数值上，但还是需要对它设定一些合理的界限。

一种可以减少样本量变化程度的方式是根据群大小来分层。在前面的例子中，我们可以将街区根据其规模分为三层：一层可能包含街区 2、3 和 6，第二层包含街区 1、7 和 8，第三层包含 4、5 和 9。从每一层中抽取一个街区可以将样本量的变化减少至 15（街区 1 或 7、6 和 9）到 31（街区 2、5 和 8）。如果选取更多单位的话，根据规模分层一般会实现对样本量的较好控制。然而，使用规模分层会减少使用其他因素进行分层。考虑到这一点，我们提出另一种控制样本量的方法，而这种方法也更为常用。

首先，设定我们希望样本能满足的条件：(1)它应该是等概率抽样样本；(2)它应该被限制在 3 个街区中；(3)样本量最好被固定在 $n=21$，无论我们抽取哪些街区。第一个和第三个条件都暗示了 $P(\alpha\beta)=1/15$。如果每一层中可以抽出 7 个元素，那么第二个和第三个条件也能被满足；在此情况下，第二阶段的在街区 α 中的选择概率为 $P(\beta \mid \alpha)=7/B_\alpha$。将 $P(\alpha\beta)=1/15$ 和 $P(\beta \mid \alpha)=7/B_\alpha$ 代入选择方程，我们就可以得到：

$$\frac{1}{15}=P(\alpha)\times\frac{7}{B_\alpha}$$

从而可以得出选取街区 α 的概率为 $P(\alpha)=B_\alpha/105$。因此，如果我们使用 PPS(Probability Proportional to Size)与 B_α 成比例进行概率抽样抽取街区的话，那么三个条件就都能够

被满足。

一般而言，对于一个使用 PPS 抽样的等概率抽样两阶段的样本选择公式如下：

$$P(\alpha\beta)=f=\frac{n}{N}=\left(\frac{aB_{\alpha}}{\sum B_{\alpha}}\right)\left(\frac{b}{B_{\alpha}}\right) \qquad [6.2]$$

其中 a 个初级抽样单位 PSU 是使用 PPS 抽取的，从每个已抽取的初级抽样单位中再抽取 b 个元素，从而得到 $n=ab$，以及 $N=\sum B_{\alpha}$。这一公式可以被延伸到三阶段抽样的情形下，其中有 a 个初级抽样单位，从每个初级抽样单位中抽取得到的 b 个二阶段单位(Second Stage Units, SSUs)，以及在每个二阶段单位 SSU 之内抽取 c 个元素：

$$P(\alpha\beta\gamma)=f=\frac{n}{N}=\left(\frac{aB_{\alpha}}{\sum B_{\alpha}}\right)\left(\frac{bB_{\alpha\beta}}{B_{\alpha}}\right)\left(\frac{c}{B_{\alpha\beta}}\right)$$

其中 $n=abc$，$B_{\alpha\beta}$ 是在初级抽样单位 PSUα 中的二阶段单位 SSUβ 的样本量，并且 $\sum_{\beta} B_{\alpha\beta}=B_{\alpha}$。

使用 PPS 选取街区可以通过将其样本量累积起来如下：

街区：	1	2	3	4	5	6	7	8	9
B_{α}：	20	100	50	15	18	43	20	36	13
累计的 B_{α}：	20	120	170	185	203	246	266	302	315

使用累计总数，每个街区就对应了一个数字：街区 1 对应 20 个数字，从 001 到 020；街区 2 对应 100 个数，从 021 到 120；街区 3 对应 50 个数字，从 121 到 170，依此类推。根据这种方式，每一个街区所对应的数字与其样本量 B_{α} 相同。一个从 001 到 315 的随机数从而可以在 PPS 下选取街区。

比如我们抽到的随机数为 197，那么就会抽到街区 5。

使用上述方法，我们可以抽取三个随机数从而得到三个街区，然而这种有放回的抽取却可能使一个街区被多次抽到。这里，人们可以使用系统抽样来实现无放回的 PPS 抽样。我们可以将总数 315 按照我们希望分组的个数，即三个，得到长度为 105 的抽样区间。一个从 001 到 105 的随机数被选取，比如 047，可以决定第一选择，即街区 2。然后，105 就可以被加到 152，使得街区 3 成为第二选择；继续此过程我们得到 257，从而得到街区 7 作为我们的第三选择。

我们也可以在 PPS 抽样中使用前面提到的末级群抽样 (unltimate clustering sampling)近似的方法。假设在 PPS 的第二阶段，我们从每个被选择的初级抽样单位中通过简单随机抽样抽取 b 个元素。然后，为了匹配末级群，我们在每一个初级抽样单位的总体中可以形成末级群，即 UC，也就是用简单随机抽样从中抽取 b 个元素形成第一个末级群，然后在剩余的元素中用简单随机抽样抽取 b 个元素形成第二个末级群，如此往复直到所有的元素都被抽取完毕。通过这种方式，包含 B_α 的初级抽样单位 α 就可以被分为 B_α/b 个末级群(我们在这里假设 B_α/b 是一个整数)。然后，在这些末级群中使用简单随机抽样就与无放回的 PPS 抽样是等价的，但唯一的不同是末级群抽样可能会从一个初级抽样单位中选择多于一个末级群，而这一点在 PPS 中是不可能的。但给定从一个初级抽样单位中抽取两个末级群的概率非常小，两种方法的差异是可以忽略的。为了看出这两种方法之间的相似点，我们需要知道在末级群抽样中，从初级抽样单位 α 中抽取一个末级群的概率与 PSU 中末级群的个数成正比，即

B_{α}/b(比如,它与初级抽样单位的规模成正比)。

因为 PPS 抽样是等概率抽样的,同时其样本量是固定的,其简单样本均值为:

$$\bar{\mathrm{y}}_{\mathrm{p}}=\sum_{\alpha}\sum_{\beta}y_{\alpha\beta}/n=\sum_{\alpha}\bar{y}_{\alpha}/\mathrm{a}$$

是一个总体均值的无偏估计量。末级群样本近似对 $\mathrm{V}(\bar{\mathrm{y}}_{\mathrm{p}})$ 给出的方差估计量(见公式 5.5),我们忽略 FPC 项可以得到:

$$\mathrm{v}(\bar{\mathrm{y}}_{\mathrm{p}})\approx \mathrm{s}_{\mathrm{a}}^{2}/\mathrm{a} \qquad [6.3]$$

另外,使用 PPS 进行第一阶段抽样和简单随机抽样进行第二阶段抽样的样本均值的近似的设计效应可以由公式 5.6 给出,即 $[1+(\mathrm{b}-1)\rho]$。

在实践中,上面描述的 PPS 抽样很少是可行的,因为我们往往不知道真实的抽样单位的规模。然而,我们却可以通过近期的人口普查或者其他数据来得到一个比较好的估计,而在其他情形下我们也可以通过其他途径来得到可靠的估计。如果我们已经有了比较好的估计的规模或者其度量,那么我们在 PPS 过程中使用它们来替代真实值一般也是可以的。然而,很重要的一点是,我们需要分辨使用真实规模和估计规模的差别;因此,我们将仅用按规模大小成比例的概率抽样(PPS)对应我们使用了真实总量的情形,而使用按估计规模大小成比例的概率抽样(PPES)来对应其他的情况。我们将估计的规模或者其度量记为 M_{α}。

与公式 6.2 相对应,在第一阶段抽取了 a 个初级抽样单位的两阶段 PPES 的选择方程是:

$$\mathrm{P}(\alpha\beta)=f=\left(\frac{\mathrm{a}M_{\alpha}}{\sum M_{\alpha}}\right)\left(\frac{\mathrm{b}}{M_{\alpha}}\right) \qquad [6.4]$$

这个公式的一个重要意义在于，为了使样本是等概率抽样的，第二阶段的抽样比率是(b/M_α)。将这一比率应用到在被选择初级抽样单位α中抽取B_α个元素，我们可以得到从该初级抽样单位中得到的期望样本量为$b(B_\alpha/M_\alpha)$。这一期望样本量会依照比率(B_α/M_α)的不同而在初级抽样单位之间有所差别，而只有当$B_\alpha=M_\alpha$时样本量为b，即当初级抽样单位估计的样本量等于其真实样本量时。为了保留样本的等概率抽样性质，我们需要接受从不同初级抽样单位中抽取的样本的变异性；如果估计的规模是足够精确的，那么这一变异性是可以接受的。

为了举例说明，我们在此假设九个街区的真实规模(B_α)是未知的。通过在这一地区快速巡游，我们会得到一些比较粗糙的估计为M_α，从而可以将其用到 PPES 中。M_α的值由下面给出，同时给出的还有每个初级抽样单位期望样本量（假设选中了该初级抽样单位）。这些期望的样本量可以通过将比率$7/M_\alpha$来代替前面的相应数字。一旦被选中，初级抽样单位的真实规模就会被决定。

街区：	1	2	3	4	5	6	7	8	9
M_α：	30	110	50	20	20	50	10	50	20
B_α：	20	100	50	15	18	43	20	36	13
期望样本量：	4.7	6.4	7.0	5.3	6.3	6.0	14.0	5.0	4.6

由于M_α并非完全准确，期望的样本量会有一些变化，但是大多数情况下，这种变化是可以接受的。但需要注意的是，如果选择了初级抽样单位 7 的话，我们会得到很大的期望样本量，这是因为初级抽样单位的真实大小（20）被低估了。在分配规模的度量的时候，我们需要注意避免低估总体

的情况，因为此时会出现一些问题。举个例子，我们根据上次人口普查估计一个街区包含10个家庭住址，然而最近新建的一栋建筑中包含了800个新的住所。另一个期望样本量的明显特征是，其大多数都小于b的值(7)。为了解释这一点，我们需要注意到 M_α 倾向于高估 B_α：$\sum M_\alpha = 360$，而 $\sum B_\alpha = 315$。为了得到一个 $n=21$ 的样本，整体的抽样比率需要被设定在21/360上；期望的总样本量因此是(21/360)315=18.4。这一差异使得人们必须试图为总体寻找一个比较好的估计。

正如我们在前面所讨论的，使用PPES抽样的一个后果是总样本量并非固定的，而是一个随机变量，其依赖于所选择的初级抽样单位。为了强调这一点，我们用 x 而不是 n 来代表总样本量，用 $r=y/x$ 来代表样本均值，其中 y 是 y 变量在样本中的总和。在这里，我们将样本均值叫做比率均值(ratio mean)或者比率估计(ratio estimator)，是因为它是两个随机变量的比值。比率均值并不是总体均值的无偏估计，但是如果 x 的变化足够小的话，其中的偏差是可以忽略的。当 x 的变化系数(coefficient of variation)小于0.1的时候，可以放心忽略这一偏差，其中变化系数被定义为 x 的标准误差与其期望的比值，即期望的样本量。

比率均值的方差比较复杂，因为它的分母是随机变量。因此，只有在大样本的情况下，基于泰勒展开(Taylor expansion)或者delta方法，我们才能对其进行近似。为了合理使用这一近似，我们需保证 x 的变化系数比较小，一般是小于0.2或者最好小于0.1。比率均值 $r=y/x$ 的近似方差估计量

的一个一般表达式如下：

$$v(r) \approx [v(y) + r^2 v(x) - 2rc(x, y)]/x^2 \qquad [6.5]$$

其中，$c(x, y)$是 x 和 y 的样本协方差。为了应用这一公式，我们需要用合适的公式来代替 $v(y)$，$v(x)$以及$c(x, y)$。为了说明这一点，考虑一个等概率抽样的分层多阶段抽样。让 $y_{h\alpha}$表示 y 变量在层 h 中初级抽样单位 α 的样本的和，让 $x_{h\alpha}$表示该初级抽样单位中的样本量，让 y_h 表示 y 变量在层 h 中的 a_h 个被抽中的初级抽样单位的和，让 x_h 表示该层的样本量。然后，使用有放回的近似，得到：

$$v(y) = \sum_h a_h s_{yh}^2$$

$$v(x) = \sum_h a_h s_{xh}^2$$

$$c(x, y) = \sum_h a_h s_{xyh}$$

其中

$$s_{yh}^2 = \sum_\alpha \left[y_{h\alpha} - \left(\frac{y_h}{a_h} \right) \right]^2 \Big/ (a_h - 1)$$

$$s_{xh}^2 = \sum_\alpha \left[x_{h\alpha} - \left(\frac{x_h}{a_h} \right) \right]^2 \Big/ (a_h - 1)$$

$$s_{xyh} = \sum_\alpha \left[x_{h\alpha} - \left(\frac{x_h}{a_h} \right) \right] \left[y_{h\alpha} - \left(\frac{y_h}{a_h} \right) \right] \Big/ (a_h - 1)$$

在此，我们需要说明公式 6.5 中 $v(r)$使用上面替换的一般性。这一公式适用于任何等概率抽样分层多阶段抽样的情况。不管抽取初级抽样单位使用什么概率以及不管在初级抽样单位内部是如何抽取子样本的，这一公式都是适用的。这一公式适用于非分层选取的初级抽样单位（特殊情况

是只有一个层)得到的样本以及样本量固定的PPS样本——此时 v(x)=0 而且 c(x, y)=0。这一条可以被应用到基于总样本和子群体(subclass,比如有工资收入的人或者已婚者)的比率均值以及百分比之上。唯一的限制在于,我们需要保证变异系数小于0.2,以及有放回的近似是合适的。通过修改 y_α 和 x_α 的定义,我们可以将公式6.5拓展到非等概率抽样的情形下。关于这一点的更多讨论,参见基什(Kish, 1965:第六章)。

在我们结束对PPS和PPES的讨论之前,我们需要指出另一个在实践中常常碰到的问题。为此,我们将前面的街区的例子做了一些修改。现在,我们需要从十个街区中抽取三个,每个街区的估计规模为 M_α,并且 $\sum M_\alpha = 315$:

街区:	1	2	3	4	5	6	7	8	9	10
M_α:	20	120	45	15	18	43	5	0	36	13

和前面一样,我们期望的样本量为21,说明子样本量应当为 b=7。在使用前面的抽样方式时,有两个问题。

首先,使用区间为105的系统抽样会导致街区2被重复选取。由于它的规模大于区间长度,因此它既有可能在样本中出现一次,也有 15/105=1/7 的概率被选取两次:如果开端的随机数在021到035之间,那么初级抽样单位2就会被同时选为第一个和第二个单位,因为给这一区间内的数加上105得到的值仍然小于140,即街区2的累计规模。一个简单的解决方案是接受这两个选择,并从这一街区中抽取两个不同的子样本。另一个解决方案是对初级抽样单位2的出现进行限制,比如说,一个层中。具体来说,我们将这一街区单

独作为一层,其中的元素使用整体的抽样比率 1∶15。然后,我们在剩余的元素中使用 PPES 抽取两个街区,其抽样比例也为 1∶15。此时,对于后面的层而言,由于 b 减少了,此时该层 $\sum M_\alpha=195$, $2b/195=1/15$, 从而 $b=6.5$。在实践中,这种初级抽样单位非常大从而可能在抽样中出现两次的情形是经常会发生的。我们经常把它们当做独立的层,将它们叫做自代表的(self-representing)初级抽样单位。

由选择方程得到的第一阶段选择概率大于 1 可以鉴别出过大的初级抽样单位;对于街区 2 而言,其被选择的概率为 $aM_\alpha/\sum M_\alpha=3\times120/315=360/315$。另一个在 PPS 或者 PPES 抽样时经常会碰到的问题是过于小的初级抽样单位,对应其第二阶段被选择的概率 b/M_α 大于 1。街区 7 就属于这一类,因为 $b=6.5$ 而其大小只有 5。处理这一问题的一个简单方法是将这一街区与地理上相邻的街区合并,然后将两个街区当做一个群。这一方法在未选之前是很容易实现的;在选择之后,人们也可以使用一些合并的规则将其合并(Kish, 1965:244—245)。如果存在很多小规模的初级抽样单位而将其合并会导致一些实地调查的困难,我们可以将它们放置到单独的层中分别抽样。一般而言,最小的初级抽样单位规模会被设定在大于 b 从而避免数据搜集时将该初级抽样单位中的元素全部抽取或者过度抽取的问题。对于 $b=6.5$, 我们可以设置最小的初级抽样单位规模为 13,从而保证子样本的抽样比率不超过 1/2。

最后,我们注意到街区 8 的 $M_\alpha=0$, 从而不可能从中进行抽样。然而,事实上 M_α 仅仅是一个估计的规模,也许其现

状已经变化，而街区 8 已经包含了一些住户。我们可以将街区 8 与其毗邻的街区合并起来，从而让街区 8 现在包含的住户有被抽到的可能性。这一方法可以避免将总体中的一些元素被抽到的概率为 0 所带来的偏误。地区抽样的一个重要特征是即使当我们已经对地区的规模已经进行了估计，每一个可以居住的地区依然会被赋予抽中的概率。

第7章

其他概率抽样设计

总结而言，前面章节中所讨论的抽样方法对于处理大部分抽样问题已经足够了。然而，我们也会讨论一些在某些情况下更加合适的其他方法，包括二阶段抽样（two-phase sampling）、重复抽样（replicated sampling）以及面板抽样设计。在这一章中，我们将讨论以上三种抽样设计。

第1节 | 二阶段抽样

在二阶段抽样或称双重抽样(double sampling)之中，人们在初期或第一期(first-phase)搜集一些信息项，然后在第二期从初期样本的子样本中获取更多的信息项。这种方法可以被拓展到多期(多期抽样，multiphase sampling)，但在大多数情况下，两期抽样就足够了。

一个使用二阶段抽样的场合是，人们对于一个调查中的不同估计值的精确度的需求是难以调和的，这就意味着我们需要不同的样本规模。在这种情况下，我们可以从第一期较大的样本中获取一些估计的信息，然后在第二期的样本中获取其他需要得到的估计。二阶段抽样不仅可以降低数据搜集和处理成本，也能够降低某些受访者的工作量。二阶段抽样的一个例子与美国的人口与住房普查有关。在近期的人口普查中，我们可以搜集到全部人口的基本人口学信息以及其他信息(第一期的样本列举了总体的全部)，然后只从总体的子样本来获得其他额外变量的信息。

人们使用二阶段抽样的另一个原因在于，抽样者希望使用某些总体的数据来进行有效的抽样设计，但从总体获取这些信息代价过大。为了更经济，可以先为了得到第一期的大样本来搜集数据，然后再从中抽取第二期的样本。使用这种

方法，第一期的样本可以提供第二阶段抽样时分层的信息，对 PPS 或 PPES 抽样时的规模的估计，或者群的信息。为了对二阶段抽样的有效性进行评估，我们需要考虑到第一期抽样的成本；因为存在这些成本，第二期样本量必须要少于单期的样本量。正由于此，只有当第一期抽样的成本比第二期抽样的成本小的时候，二阶段抽样才是有用的。当我们使用不同的方法在两个阶段中搜集数据时，两期的成本可能存在很大差异：比如在第一期，我们使用邮件或者电话访问，而在第二期中使用面对面或者更昂贵的访问形式（比如一些医疗调查）。

在调查少数人群的时候常常使用二阶段抽样，也就是说，总体中的子群体没有明显的抽样框：比如越南退伍军人、黑人，或者近期退休的人。对少数人群进行经济而有效的抽样设计是抽样调查者常常面临的一个富有挑战性的问题（Kish, 1965：第 11.4 部分）。一个可以考虑的方法是在二阶段抽样的第一期样本中将少数人群的成员以不高的成本标记出来，然后在第二期中对他们进行更深入的调查。本质上，这一方法采取的是两期分层抽样的步骤。根据其是否属于少数人群，第一期样本的成员被分为两个（或更多）层中。然后，人们对不同的层不按比例进行抽样。如果第一期对少数人群的标记是没有误差的，那么我们可以设定少数人群的成员的层抽样比率为 1，而非成员的层的抽样比率为 0。然而，如果这一标记是有误差的，那么对于第二层的抽样比率需要被设定为非 0，从而让被错误地分配到该层的少数人群有非零的抽中概率。当第一期的标记并不完美而人们一定会犯错时，多报比少报要好一些，因为前者更加容易处理。

比如，在一个研究听力严重受损的儿童的调查中，最初的筛选可以使用对听力受损的比较松弛的定义，从而保证第二期的研究中包括所有听力严重受损的儿童，第二期听力受损程度可以在可控的实验室条件下来测量。

为了说明二阶段抽样在群中的应用，我们以一个在欧洲城市中进行的对选民政治观点的调查为例。我们将该城市所有选民按字母顺序排列的包含居住地址的名单作为抽样框。由于该城市规模很大，而调查方式被设计为面对面进行，因此我们希望在抽样中使用群来减少访问员的旅途成本。从理论上讲，选民的地址可以被用到对群的分配中，但是其成本却是异乎寻常地高。相反地，我们可以选取一个十倍于要求样本的样本，然后将其基于地理位置的相似性分为等规模的群，最后从中抽取 1/10 作为最终的样本。

第2节 | 重复抽样

在重复抽样或贯穿抽样法(interpenetrating sampling)中,总样本由一系列重复抽取的子样本构成,每一个子样本都是使用同样的抽样方法得到的。重复抽样可以用来研究变量的非抽样误差(nonsampling errors),比如由不同访问员和编程者得到的结果的变动以及辅助计算变量的标准误。这种方法的精髓在于,每一个子样本都能够提供独立的、可比的、对总体参数的估计。

在此,我们举一个为了研究访问员效应(interviewer effects)而进行重复抽样的例子,其中要求使用简单随机抽样抽取容量为1 000的样本,由20个访问员完成任务。在无重复抽样的情况下,可能会根据地理的便利程度,将1 000个受访者分配给不同的访问员,比如将最难采访的地区的受访者分配给最优秀的访问员。当一个访问员不能成功地采访被访者时,人们也许会派遣另一个更有经验的访问员重新采访。由于这一对访问员的分配并不是随机的,访问的结果当然也会根据访问员不同而出现差别,但由于这一差别同样可能是由样本中受访者的差别造成的,我们无法厘清这两种差别的来源。

在一个简单的重复抽样设计中,我们使用20次独立的

简单随机抽样抽取出规模为1 000的总样本，每个子样本的规模为50，然后每个访问员在一个子样本中进行50次访问。由于这些样本是完全可比的，任何除去抽样波动的子样本的差别都可以被归结为来自访问员之间的系统性差别。具体来说，我们可以使用单向的方差分析(analysis of variance，参见Iversen and Norpoth，1976)来区分抽样的波动与真实的差异；然而，当重复抽样的过程采取比较复杂的抽样设计时，对应的计算方式也会不同。

为了描述访问员变化(interviewer variance)的计算方式，我们令 $\bar{y}_1$，$\bar{y}_2$，…，$\bar{y}_c$ 表示从c个子样本中得到的均值，其中每个子样本被分配到一个访问员下。这些c个均值的方差可以根据 $v_1=\sum(\bar{y}_\gamma-\bar{y})^2/(c-1)$ 进行估计，其中 $\bar{y}=\sum\bar{y}_\gamma/c$ 是子样本均值的均值。这一估计量对于是否存在系统的访问员效应没有任何假设；当这一效应存在时，我们期望这一估计量大于不存在对应的情形。在不存在访问员效应的零假设下，我们可以使用简单随机抽样理论来提供对于 $\bar{y}_\gamma$ 的方差的另一个估计量：在公式2.3中，忽略FPC项，得到 $v(\bar{y}_\gamma)=s_\gamma^2/r$，其中 s_γ^2 是在第 γ 个子样本中的元素的估计方差，而 $r=n/c$ 是子样本量。在c个子样本中的 $v(\bar{y}_\gamma)$ 的估计量的均值为 $v_2=\bar{s}^2/r$，其中 $\bar{s}^2=\sum s_\gamma^2/c$ 为在子样本内的方差估计量的平均。比较 v_1 和 v_2 我们可以得到一个对零假设的检验。这一比较可以通过取得二者比值 $F=v_1/v_2=rv_1/\bar{s}^2$ 获得，其中比较大的F值就说明存在访问员效应。对F大于1的显著性检验可以通过标准F检验进行，其中 $(c-1)$ 和 $c(r-1)=(n-c)$ 为其自由度。一个有用的标记

访问员变化的指标是层内相关性系数(intraclass correlation coefficient)ρ,我们用其度量 y 值的总方差中可以被纳入访问员变异的部分。对 ρ 的估计可以由 $(F-1)/(F-1+r)$ 得到。对于具体例子,请参考基什的研究(Kish, 1962)。

访问员中存在的变动得到的结果与群的结果比较相似,每一个访问员的分配其实是单独的整群效应。因此,与整群样本的设计效应相同,重复抽样中的访问员变动的效应可以由将简单随机抽样的总体样本均值的方差与 $[1+(r-1)\rho]$ 得到。对于整群抽样的情况,即使是一个很小的 ρ 也可能会导致很大的乘数效应,因为 r,即每个访问员进行的访问次数可能非常大。根据简单随机抽样理论得出的对整体均值的方差的估计(公式 2.3)并不允许整群效应或者访问员效应的出现。事实上,基于子样本之间的变化使用重复抽样得到抽样方差的估计量是很有吸引力的,因为它能够自动包含访问员变动的整群效应。正如我们下面要说的,这一方差估计量事实上为 v_1/c,是简单整群抽样方差的估计量(公式 5.4)s_a^2/a 的另一种表达形式。

使用重复抽样来研究系统性的访问员效应或访问员变化的成本的一个重要来源是,我们需要随机选择访问员,而不是以更有效率的安排为目标。而研究访问员变异的便利程度则依赖于一般的调查条件;比如,我们更容易在电话访问而不是面对面访问中实现这一点,而对于面对面访问来说,更容易实现小的、紧凑的样本。对于一个非常分散的、总体的多阶段抽样,如果我们完全随机地分配访问员,那么访问员的旅途成本将明显增加,但完全的随机分配一般并不是必须的。如果我们采取某些有限形式的重复抽样,比如在某

一初级抽样单位或者层中使用随机访员分配，仍然能够让人们对访问员变异进行估计。

使用重复抽样的另一个原因在于，人们可以提供对简单方差的估计。给定我们有 c 个从独立重复抽样中得到的对参数 Z 的估计，z_1，z_2，…，z_c，其均值 $\bar{z}=\sum \bar{z}_\gamma/c$ 的方差的估计可以由下式给出：

$$V(\bar{z})=V(z_\gamma)/c$$

其中 $V(z_\gamma)$可以由 c 个值估计得到：

$$v_1=\sum(\bar{z}_\gamma-\bar{z})^2/(c-1)$$

从而

$$v(\bar{z})=v_1/c=\sum(\bar{z}_\gamma-\bar{z})^2/c(c-1) \qquad [7.1]$$

可以给出一个根据重复抽样估计方差的一般形式。这一公式可以被应用到任何形式的统计中（比如指数、相关与回归系数，以及简单均值和百分比），另外子样本设计也可以采取任何复杂的形式（比如分层多阶段 PPS 设计）。

使用公式 7.1 的一个小问题在于，它给出了重复值 $\bar{z}$ 的均值的方差。这一均值一般与将子样本合并成为一个大样本得到的估计量 $\tilde{z}$ 并不相同，而事实上 $\tilde{z}$ 则是我们更加偏好的估计量。然而，在实际中两者的区别一般都很小。人们通常采用的方法是，计算 $\tilde{z}$ 并且使用公式 7.1 或者其简单变换来计算 $v(\bar{z})$，从而提供一个对 $\tilde{z}$ 的方差的估计。

使用这一方法的一个重要方面在于对 c 的选择，即使用多少重复样本。如果我们选择很小的 c，对 $v(\bar{z})$的估计则会不精确，而这一问题会影响到对我们感兴趣的参数构建置信

区间的宽度。给定我们有 c 个重复样本，$v(\bar{z})$有$(c-1)$个自由度；因此，在构建置信区间的时候，就可以使用有$(c-1)$个自由度的 t 分布。为了说明这一效应，让我们考虑一个有1 000 个元素的简单随机抽样。使用通常的方法，我们得到$\bar{Y}$的 95%的置信区间为 $\bar{y} \pm 1.96s/\sqrt{n}$，其中 1.96 是从标准正态分布表中得到的。如果我们有容量为 100 的 10 个子样本的重复抽样设计，其 95%的置信区间为 $\bar{y} \pm 2.26/\sqrt{v_1/10}$，其中 2.26 是从具有 9 个自由度的 t 分布表中得到的。如果我们有容量为 250 的 4 个子样本的重复抽样设计，其 95%的置信区间为 $\bar{y} \pm 3.18/\sqrt{v_1/4}$，其中 3.18 是从具有 3 个自由度的 t 分布表中得到的。由于在每种情形下，标准误差的估计量都是对真值无偏的，包含了 10 个子样本的重复抽样的估计得到的置信区间比使用通常估计量大 15%，而包含了 4 个子样本的设计得到的估计的置信区间则会比通常的估计量大 62%。为了得到一个比较合理的精确的方差估计量，我们需要比较大的 c，从而进行更少的分层。这一情况的出现是因为每个子样本必须至少从每一层中选择一次。在多阶段设计的情况下，对分层的限制造成的危害尤其大。为了说明这一点，我们选取了 60 个初级抽样单位。在通常的方法下，初级抽样单位应当被分为 60 个层，其中每层包含一个选择，或者分为 30 层，其中每层包含两个选择。如果我们有一个包含 10 个子样本的重复抽样设计，那么层数的最大值则被减少到了 6。

总体而言，使用重复抽样得到的简单的方差估计量的代价是对精度的损失：如果 c 很小的话，那么方差估计量的精度则会由于自由度的限制而减少；如果 c 比较大，那么调查

估计量本身则会损失分层的层数。由于这些原因，简单重复抽样实际上并没有得到广泛应用。相反，人们发展出了一些伪重复抽样技术(pseudoreplication techniques)来使得分层仍然能够被使用，并同时给出可观的精度。我们将在第10章中讨论这些技术。

第3节 | 面板设计

在前面的章节中，我们暗自假设了抽取的样本是截面的(cross-sectional)，从而只进行了一轮数据搜集。然而，事实上有很多调查的目的要求在两个或者多个时点上进行数据搜集。虽然前面介绍的抽样方法在此依然适用，但是由于现在考虑到时间维度，我们需要进行一些抽样方法的说明。

进行多轮的数据搜集的一个目标是捕捉信息随时间的变化。在这里，我们需要区别总变化(gross changes)和净变化(net changes)，前者指的是在元素级别的变化，而后者指的是在加总层面的变化。如果我们需要个人层面变化的度量，比如在研究休闲活动的变化对血压的影响时，那么我们就需要对同一样本进行多轮的数据搜集。只要我们关心的是净变化，比如说研究某一政治领袖的受欢迎程度变化，那么我们并不需要对相同的样本进行追踪。但是，即使对于净变化而言，人们对于相同的样本进行追踪往往是更有效率的。

在不同时点进行访问的另一个目的在于，人们可以在数据可得或者能够被准确汇报的时候对信息进行采集。因此，当我们在调研中希望记录家庭年收入时，就可以在一年中的不同时点进行采访，从而可以在人们的记忆比较清晰时采集信息。另外，在一个研究儿童学前教育与学校表现的研究

中，人们几乎必须要在学前以及就学两个时点进行数据搜集。仅仅依赖于事后人们对此的回忆是不保险的，因为人们可能会因为在学校的表现不同而扭曲了学前阶段的回忆。

在面板研究或纵贯研究(longitudinal survey)中，人们需要在不同时点对相同的个体进行访问，而这一点会引发一些在横截面研究时不会出现的问题。一个问题在于被调查者的迁移。在众多面板研究中，很多元素，包括个人或者家庭，会在面板研究的时间段内迁移。这些迁移者需要被保留在面板研究中，因为我们需要保持在开始时选择的概率样本的完整性，这一点也要求我们发展出更有效的追踪方式。由于一些迁移者会离开在多阶段抽样中抽取到的初级抽样单位，这会使得之后几轮的面对面访问的数据搜集成本大幅增加。

面板研究的第二个问题在于总体会随时间发生变化，一些在初始总体中的元素不再属于总体，同时也会有一些新元素加入。在此，我们需要考虑一个长期的关于某社区健康的面板调查。起初，我们抽取了这一社区成员的概率样本，然后对他们进行数年的追踪。在这一时间段内，该社区的人口可能会发生变化：一些原始的居民会离开，比如有些人去世或者有些人搬离了该社区，同时也会有一些新的居民加入，比如新生儿的出生以及搬入社区的人。离开社区的这部分人会导致样本量的减少，但该面板仍然是该社区的原始居民中没有离开的人的一个概率样本。而新进入社区的这部分人则导致他们在样本中没有被代表。因此，这一样本对于所有该社区的总体来说并非一个概率样本，因为该总体的结构在变化。当一个总体中有相当比重的新进入者，而我们需要之后出现的总体对应的横截面的结果时，我们需要一个新进

入者的补充样本(supplement sample)。如果我们研究的元素是一个组时,如住户,那么问题会比较复杂。大量的住户在短期之内(比如一年)的构成可能会变化,从而产生面板研究中的一些概念和实际操作的问题。

面板研究的另一个问题在于反复采访可能会对受访者产生负面影响。一些人可能因为负担太重而拒绝采访,甚至拒绝继续留在面板里,从而导致面板中成员的偏误(参见第9章关于无应答的介绍)。而一些人可能因为在面板中从而对调查的主题有一定了解,进而导致他们给出一些不典型的应答。这种面板效应可能会出现在一个访问消费者的面板调查中,其中要求受访者定期汇报他们的家庭购买行为。这种定期汇报的形式可能会使被调查者对价格更加敏感,从而改变他们的购买习惯。另一个与此相关的面板研究中的问题在于,受访者可能对他们之前的回答有记忆,从而试图给出前后一致的回答。

一个常用的用来解决面板调查中出现的以上问题的方式是限制元素在面板中持续的时间,具体办法是采用面板轮换的方式(panel rotation)。举个简单的例子,面板中的每一个成员可能在为期三轮的调查中都被留下。在每一轮中,1/3的前一轮的样本会被扔掉,同时新的1/3的样本会被加入:这些新人会被包含到接下来的两轮中。因此,使用字母来表示样本的三个部分,如果第一轮的样本为ABC,第二轮为BCD,第三轮为CDE,第四轮为DEF,依此类推。在这种方式下,相邻的两轮的样本存在2/3的重叠,而隔一轮的两个样本则存在1/3的重叠。

正如我们之前观察到的,面板设计可能是很有用的,但

是对于净变化而言并不是必须的。让我们考虑对时点1和时点2的变量 y 的均值的变化的估计量 $\bar{y}_2-\bar{y}_1$。这一差别的方差由如下公式给出：

$$V(\bar{y}_2-\bar{y}_1)=V(\bar{y}_1)+V(\bar{y}_2)-2\bar{R}\sqrt{V(\bar{y}_1)V(\bar{y}_2)} \quad [7.2]$$

其中 $\bar{R}$ 为样本均值 $\bar{y}_1$ 和 $\bar{y}_2$ 之间的积矩相关系数(product-moment correlation coefficient)。如果我们有两轮不相关的样本，那么 $\bar{R}=0$。如果两轮样本间有重叠，那么 $\bar{R}$ 不为0；它一般是正值，但有时也是负数。公式7.2的最后一项反映了人们在使用面板设计时对估计量的精度的收益（正的 $\bar{R}$）和损失（负的 $\bar{R}$）。

为了更加深入地理解样本重叠对于测量变化的度量的影响，我们现在考虑一个简单情形，其中包含静态的总体以及样本量为 n 的简单随机抽样；另外，假设（一般而言是可以接受的近似）两期的元素方差是相等的（即 $S_1^2=S_2^2=S^2$），在此我们忽略FPC项。接下来，如果两个样本有比例为P的重叠，公式7.2就可以简化为：

$$V(\bar{y}_2-\bar{y}_1)=2S^2(1-PR)/n$$

其中R是元素 y 的值在两期的相关系数。有两种特殊情况：一是两个独立的样本对应的 $P=0$，第二种就是完全重叠的两个样本的 $P=1$。当 $P=0$ 时，这一差别的方差为 $2S^2/n$，因此在面板研究中的这一差别的方差与两个独立样本的方差之比为 $(1-PR)$。为了继续说明，假设个人的政治态度（或者血压）在两期的相关性为0.75。那么，完全重合的面板会将 $\bar{y}_2-\bar{y}_1$ 的方差以乘数 $(1-0.75)=0.25$ 减少。2/3的重

合(1/3 被替换掉)会将 $\bar{y}_2-\bar{y}_1$ 的方差以乘数[1−(0.75×2/3)]=0.5 减少。如果跨时的相关性很高,那么面板设计的优势在衡量变化方面的作用是很大的。在轮换面板设计的情形下,人们可以使用更复杂的估计量来获得更多的收益(参见 Kish,1965:463—464)。然而,需要注意的是,如果 R 是负数(比如当 y 变量表示在上个月的耐久性商品的购买),那么面板设计会导致在衡量变化方面的精度的损失。比如,当 R=−0.2时,以及在完全重叠 P=1 的情形下,(1−PR)=1.2,使得这一变化量的方差比两个独立样本的方差大 20%。

最后需要说明的是,在公式 7.2 从非独立样本中的正相关中得到的收益并不局限于相同的元素都在面板中的情形。尽管相关系数 $\bar{R}$ 一般会小于相同的元素被保留的情形,保留相同的群但是选择不同的元素的抽样设计对于度量变化也是有益处的。一个有效避免追踪迁移者的设计是对住址而不是住户进行抽样,因为一个离开某一住址的住户会被新来的住户所取代。



抽样框

抽样框是抽样调查中的一个重要组成部分。它不仅提供了一个识别和定位总体中的元素的方式，而且经常包含很多额外的可以用来分层或者聚类的方式。抽样框的组织也经常对抽样设计发挥重要作用。比如，地区聚类的实现在很大程度上依赖于一个合理安排的地理单位的框，而且分层的实现也依赖于根据一些分层变量构成的群体的框。人们一般将常用的抽样框存储到电脑中，从而便于重新安排来满足抽样的需求。

一个理想的抽样框需要将总体中的每一个元素有且只有一次地列出来，并且不包含其他排列。在实践中，很难实现这一理想，而抽样调查者需要了解它们的不完美之处。基什(Kish，1965:53—59)提供了对潜在抽样框问题和解决方案的四重分类。这四个问题分别是：

——缺失元素(missing elements)：即总体中的某些元素不被包含在抽样框中；

——群：某些列举是对于元素的组而言，而非元素本身；

——空白或者外来元素(foreign elements)：某些列举并不与抽样调查的总体中的元素相关；

——重复列举(duplicate listings)：一些总体中的元素不止一次被列举。

下面，我们将讨论这些问题以及其解决方案。

第1节 缺失元素

在前面的学生调查中,假设我们拿到了在校学生的名单。关于这一抽样框的第一个问题,就是它是否包含了我们目标总体的全部学生。如果抽样框是不够的(inadequate),即该抽样框目标不是包含总体,或者该抽样框是不完整的(incomplete),即它没有包含应该包含的总体中的一些元素,那么在这两种情况下就会出现缺失元素问题。不够与不完整抽样框之间的差别在现实中非常重要,因为前一类更加容易识别。比如,学校名单如果刻意排除了总体中的非全日制学生部分,那么它就是不够的;如果学校名单由于过时而没有包含一些新的学生,那么它就是不完整的。

存在缺失元素是抽样框最严重的问题,因为除非人们找到一个方法来补救,这些元素永远不可能被抽取,样本也因此失去了对总体的代表性。有时可以绕过这一问题,方法是通过定义将这些缺失元素排除在抽样调查的总体之外。这一方法虽然不完美,但是如果排除的元素是总体中可以被忽略的一小部分,这一排除对抽样调查对象仅仅有很小的影响,同时在没有其他处理方法的情况下,可以采用这个排除的方法。一个更好的方法是寻找补充性的抽样框来覆盖缺失元素,比如,使用特殊学生以及新生的名单。这一方式可

能产生重复的问题,因为一些元素可能不止一次出现在名单上,但这一问题可以通过下面的方法解决。

通常来说,人们并不能得到合适的补充抽样框来覆盖缺失元素,因此可以寻求一个包含某种形式的链接程序(linking procedure)的方案。链接程序的目的是将缺失元素以一种被清晰定义的方式附加到特定的名单中。当人们选定了一个名单,它的元素以及任何与其链接的缺失元素就被当做群来抽样。因此,链接会产生群的问题,而这一问题可以通过下面描述的方法来解决。在我们前面所举的学校的例子中,假如抽样框包含了字母序列的班级最初入学的学生名单。一个将缺失学生链接的方式则可以是定义每个字母序列的名单包含被列出的学生以及班级中缺失的学生,其中缺失的学生出现在该被列出的学生之后并在下一个被列出的学生之前。为了在名单开始前将缺失学生覆盖,这一名单可以被当做循环的(circular)。因此,任何一个处在名单中最后一个被列出的学生后面的或者在名单中第一个被列出的学生之前的缺失学生就被链接到名单中最后一名学生。这种形式的链接是半开区间(half-open interval)的一个例子,而这一方法可以被用到很多其他场合中。一个广为人知的应用是以街道顺序在住址名单中抽取住址,而街道的每一边都被当做独立的,使用这种半开区间,缺失元素可以被链接到名单中最后被列出的住址之上。

第2节 | 群

正如之前指出的，使用链接的方法会产生抽样框中的元素的群的问题。然而，群的问题也可能在其他场合出现——比如，当我们希望对个人或者住户进行抽样，然而抽样框是住所的时候。一个解决方法是将被抽取的群中的所有元素都包含进去。这一方案的好处在于能够使元素在列表中与在样本中以同样的概率出现；尤其当列表是按照等概率抽样进行抽样的，那么元素也是按照等概率抽样抽取的。当元素为家庭而群为住址的时候，这一方案仍然是可行的，理由如下：首先，大多数住址仅包含一个家庭，包含一个以上家庭的住址很少；另外，在同一住址采访多于一户很少会有实地问题。

一个方案是从全部群中进行抽样，正如我们前面讨论过的，当群的平均规模比较大而群内相关性比较高时，整群抽样会带来比较大的设计效应。如果设计效应非常大，我们可以在群中再进行抽样来缩小群的规模。对于某些类型的群而言，再次抽样的另一个原因在于担心在某一个群中可能出现应答污染（contamination of responses）。考虑到这一问题，人们一般可以在每个被抽取的群中再抽取一个元素。这一问题在态度调查时较常出现，其中个人为元素而群为家庭

(或者住址):当一个家庭中的两个或者更多的人接受采访时,后面的受访者可能会受到之前采访的干扰。后面的受访者也可能更加不愿意合作,甚至也许会拒绝回答其中的某些问题,因为他们已经从前面受访者的经历了解到了问卷的内容和长度。当我们从一个包含 B_α 群中抽取单个元素时,每个元素被抽取的概率为$(1/B_\alpha)\mathrm{P}(\alpha)$,其中 $\mathrm{P}(\alpha)$为该群被抽取的概率。如果群是按照等概率抽样抽取的,那么元素的样本就是非等概率抽样的,因此我们需要在研究中调整相应权重(见第 10 章)。

为了避免选择偏差,从已被选取的群中抽取元素必须严格按照概率机制实现。我们从一个家庭中选取符合条件的成员中的一个作为应答者进行面对面访问。此时,访问员最好在第一次采访此住户的时候以随机抽样的形式来选取受访者,如果选取的受访者在家的话可以随即完成访问。一个可能的方法是,访问员首先将所有符合条件的家庭成员列成表,然后使用随机数表来选取一位成员。这一方法的一个严重弱点在于,出了问题检查不出来,访问员有时可能选择在家的并且较为配合的受访者,而没有正确使用这一方法。

一种广为使用的用来选取住户中的受访者的替代方法被称为基什表选择法(the Kish selection grid)。这一客观并可以检查的方法是,访问员以一种清晰定义的顺序记录符合条件的住户成员,并将其填入问卷中包含的带数字编号的表格中,然后访问员从中读出被选的受访者的编号。一个比较方便但不失准确性的给住户成员排序的方式是将他们按照性别和年龄排序。因为排序中我们仅仅需要相对年龄,所以很少需要询问其绝对年龄,代的差异一般在不同性别的年龄

序中已经足够。为了给这一方法举出简单的例子，假设我们需要调查有工资收入者，以及一个家庭往往包含不超过四个这样的成员。在一个特定的问卷中，一个指示调查员选取访问哪一个受访者的表格如下：

如果该家庭中具有工资收入者的人数为：	1	2	3	4
访问编号如下的工资收入者：	1	2	2	3

第二行的数据根据问卷的不同，参照表 8.1 而不同。

在一个只包含一个工资收入成员的家庭中，该收入者将会被选择。在包含两个工资收入者的家庭中，如果问卷包含了表格 A，B 或者 C 时，第一个被列出的工资收入者会被采访；如果问卷包含了表格 D，E 或者 F 时，第二个工资收入者会被采访。表 8.1 的第二栏给出了问卷中包含某一个表格的比重，从中我们发现问卷包含表格 A，B 或者 C 的比例是 1/2，而问卷包含表格 D，E 或 F 的比例也是 1/2。因此，在一个包含两个工资收入者的家庭中，每个工资收入者都是有可能被选入样本的。同理，在一个包含三个工资收入者的家庭中，每个成员被选中的可能性都是相等的，而在一个包含四个工资收入者的家庭中，每位成员被选中的可能性也是相等的。因此，尽管从一家庭中选择一个工资收入者将导致工资收入者在有不同数目的工资收入者的家庭被选择的概率不同，基什表选择法却可以让我们在给定家庭中的所有工资收入者的被选概率是相等的。

正如这里描述的，这一程序假设一个家庭最多包含四个工资收入者。当我们提高这一上限时，就需要更大的表格。当我们从美国的家庭中抽取成人时，一般设置人数上限为六个。基什（Kish，1965:399）给出了当人数上限为 6 时所用到

的八个表。这些表格使得规模为1，2，3，4和6的家庭成员被选择具有相同的概率，然而规模为5的家庭成员的被选概率却是不完全相等的。在包含六人以上的少数家庭中，一些成员可能不能被代表。然而，这一缺漏不是大问题，在实际中并不重要。

如果人们通常采用等概率抽样抽取群，第一种解决群的问题的方法是将所选取的群的所有元素包含在内，从而得到一个元素的等概率抽样样本。然而，这一方案一般是不可接受的，因为它很有可能被污染。第二种方案是在所选取的群中随机抽取一个元素，这可以避免样本被污染的风险，却可能改变抽样的概率。第三种方案使用了两期抽样方法，在每一群中抽取一个元素然后保留一个等概率抽样样本。使用这种方法，人们在第一期的抽样中抽取群并且在群中列出其中的元素。然后，第二期的抽样从前面的列表中抽取元素。举例而言，假设群为住户，并且家庭规模在六人以下。第一期中我们抽取住户样本，从而产生一个所需要成人数量的六倍的名单。然后，从这一名单中使用系统抽样抽取要求数目的成年人，每个住户中抽取的成年人个数不超过一个。

第3节 | 空白与外来元素

空白元素和外来元素指的是已经不在总体中的元素(比如去世者、迁移者或者已经被抹去的住址)或者虽然在抽样框中但并非抽样调查所关注的元素(比如在对工资收入者调研时未被雇用的人们)。为了简单起见,我们用"空白"(blanks)简称空白与外来元素。

处理空白元素的方法非常直接:人们只需要在抽取到该元素时将其忽略即可。这一方法已经在我们前面高中的例子中有所体现,其中一些已经离开学校的学生成为了空白。空白对于抽样框的主要影响是样本量小于我们选择的数量,因为我们会抽到一些空白并且扔掉它们。在决定抽样比率以及理想样本量时,人们需要将这一点铭记于心。在系统抽样中常犯的一个错误是使用空白元素的下一个元素进行代替。人们应当避免采取这一方法,因为它增加了下一个元素被选择的概率:该元素可能会被直接选取或者因为之前的空白元素被选取。在系统抽样下,抽样区间应当在总体中被重复,而其中空白元素可以从样本中被剔除。

在实际操作中,区分什么情况下能够在抽样框中识别空白元素很重要。如果能识别,当它们被抽中的时候,人们将其删去即可,然而在无法识别的情形下,需要在删除它们之

前与其取得联系。举例而言，在一个男人和女人的列表中抽取男人的样本时，人们会在选择阶段通过人们的名字来去掉几乎所有女人；然而，对于一个对 40 岁至 64 岁人群的调查，人们则需要采取筛选访问(screening interviews)来决定被抽取的个人是否可以被包含到调查中。在对少数群体进行抽样调查时，一个困难在于调查的目标总体仅包含抽样框很小的一部分，并且抽样框并未提供识别少数族群的方式。正如上面所提到的，一个对少数族群进行识别的方法是使用两阶段抽样，其中在第一阶段使用相对经济的抽样过程来识别少数族群。

第4节 重复列举

当抽样框由数个列表组成时，重复列举（duplicate listings）的问题会经常出现，因为一些元素可能会在多于一个列表中出现。当元素为小组时，比如家庭，而当列表为个人时，这一问题也会出现。重复元素带来的问题在于，元素被抽取的概率随着它们被列举的次数而变化。一个可能的解决方案是在总的抽样框中将重复列举去掉，然而这常常并不可行。第二个可能的方案是使用独特识别（unique identification），即将每个元素与其中某一个列举以一种被清晰定义的方式联系起来（比如，第一个列举或者最旧的列举），然后将该元素的其他列举作为空白元素处理。应用这一方法的一个案例是英国的选举人登记中的家庭抽样。在城市地区，选举人被编号并且在以街道地址划分的投票区被列出。然后，根据选举编号，人们可以通过系统抽样得到一个选举人的样本。如果被选取的选举人是在该地址被列出的第一个，那么该地址则被选取，而该地址的第二个或者接下来的选举人则被当做空白元素。然后，人们可以采取通用的方法来解决群的问题，所有该地址的家庭都被包含在了样本内。

有时候，抽样框的组织或者其包含的信息并不能够轻易让人们使用独特识别的方法。在这种情况下，独特识别可以

被应用到实地调查中，让受访者提供他们的列表信息。然而，一般而言，联系受访者的过程将成为调查费用的很大一部分，因此将一些元素作为空白将其删去的做法是不经济的。一个替代性的方法是接受所有的选择，同时在分析中使用加权的方式来调整元素不同的选择概率（参见第10章）。

第9章

无应答

概率抽样通过将抽样框中的每个元素赋予一个已知且非零的被选择的概率来避免选择偏差(selection bias)。上一部分中,我们介绍了用来消除或者降低由于抽样框不完美所导致的问题的方法。给定一个好的抽样框,我们就可以从总体中抽取一个概率样本,然而在实际搜集数据的过程中,我们仍然会遇到一些问题。无应答(nonresponse)或者说在抽样调查中不能搜集到某些被选择元素的信息,就是近年来我们在抽样调查中常常遇到的问题,因为公众现在越来越不愿意参与调查了(Steeh, 1981)。

无应答带来的潜在问题是,无应答者与应答者可能在我们的调查变量方面存在系统性差别,因此根据抽样调查得到的估计量会与基于总体估计的参数有偏差。为了更深入地理解无应答带来的偏差,我们来考虑一个简单的模型,其中总体被分为两组——应答者和无应答者;我们可以将这两组人想象成应答者和无应答者的层。然而,在现实中受访者是否提供应答也有运气成分,因此这一模型实际上过于简化,但这一简化模型对于我们的分析目的而言已经足够了。另外,为了简单起见,我们假设抽样调查需要对总体的全部元素进行完整编码。假设调查的目的是得到 $\bar{Y}$,即总体均值。

这一均值可以被表示为：

$$\bar{Y} = W_r \bar{Y}_r + W_m \bar{Y}_m$$

其中 $\bar{Y}_r$ 和 $\bar{Y}_m$ 分别为应答者和无应答者层的均值(下标 r 代表应答者，m 代表缺失，即无应答者)，W_r 和 W_m 则为两群体在总体中所占的比重($W_r + W_r = 1$)。因为抽样调查无法得到无应答者的信息，我们只能得到 $\bar{Y}_r$ 的估计。而 $\bar{Y}_r$ 与总体参数 $\bar{Y}$ 的差别为：

$$\begin{aligned}\bar{Y}_r - \bar{Y} &= \bar{Y}_r - (W_r \bar{Y}_r + W_m \bar{Y}_m) = \bar{Y}_r (1 - W_r) - W_m \bar{Y}_m \\ &= W_m (\bar{Y}_r - \bar{Y}_m) \qquad [9.1]\end{aligned}$$

这一差别，即当我们使用应答者均值替代总体均值时，依赖于两个因素：一是 W_m，即总体中无应答者的比例；二是 $(\bar{Y}_r - \bar{Y}_m)$，即应答者与无应答者的均值差。如果应答者和无应答者的层是随机形成的，那么二者的期望均值应当相等，因此我们的估计不存在偏差。然而，在实际中，我们假设无应答者为随机是非常危险的；事实上，我们通常有一些理由证明这一表现并非随机。因此，唯一能够确保无应答偏差在较小范围内的方法是使得无应答的层足够地小，从而使得 $(\bar{Y}_r - \bar{Y}_m)$ 与 W_m 的乘积不至于太大。因此，抽样调查者往往需要费尽心思降低无应答率。

在讨论无应答问题时，区分其可能发生的两个层次是很有用的：总无应答(total[or unit] nonresponse)，即我们没有从该被抽取元素上获得任何信息，以及项目无应答(item nonresponse)，即有一些信息没有从该被抽取元素上获得。总无应答常常被简称为“无应答”。接下来，我们会依次介绍总无应答以及项目无应答。

在面访调查中，总无应答可以被分为以下几类：拒绝被访；无法联系到受访者（如不在家或找不到）；受访者由于疾病、耳聋或者语言不通，从而无法进行访问；甚至问卷丢失的情形也被包含在内。在以上几类中，拒访以及不在家占据主要地位，而其他因素在大多数抽样调查中都不是很重要。在信件调查（mail survey）中，一些样本中的人们也许会发送一个他们不希望参加访问的回执，而有时候他们的邻居或者亲属会回复说受访者生病无法提供答复，有时一些问卷会因为查无地址而被邮局退回来。然而，大多数信件调查的无应答仅仅是问卷并没有被寄回来。导致这一结果的原因有很多，包括受访者拒绝回答，无法完成问卷或者问卷无法被送达到受访者处。

为了尽可能地降低拒访数量，人们在抽样调查中使用了很多方法，甚至数据搜集模式的选择也往往受到拒访的相对风险的影响。在面访调查中，访问员受到细致的培训以尽可能避免拒访，他们往往会询问受访者一个更加方便的时间对其进行访问。访问员往往会对受访者强调该访问的重要性，并常常提及该访问由一个声誉卓著的赞助机构支持；一个好的赞助机构在信件调查时尤为重要。另外，访问通常会强调调查的匿名性和保密性，从而消除受访者对其应答会被用做他途的担心。问卷开始往往会安排简单而无威胁性的问题，从而避免受访者看到问题感到尴尬或者担心需要上税而终止调查的风险。拒访率变化很大，常常依赖于问卷调查的内容、问卷长度以及调查团队的技术。

人们一般使用回访来解决拒访中不在家的问题。在面访中，当访问员无法联系到受访者时，他们一般会被要求至

少回访四次，并且四次回访需要在不同的时间（日期以及一天中不同时段，甚至晚上）进行。如果访问员在某个不在家受访者的周边地区进行访问，那么他们甚至被鼓励做第五次回访。在提高联系受访者成功率方面，预约是一个有用的办法。在电话访问中，用到回访的次数更多；因此电话访问中打电话的数量显然要比面对面访问打电话的数量多。在信件调查中，与回访类似的一种做法是跟进（follow-up），即给无应答者再次寄去访问的请求。然而，跟进的做法并不是为了解决不在家的问题，而是为了提高应答率。一个常常被用到的方法是给在一定时间段内未寄回信件者发送提醒信件，并且给上一阶段没有提供应答的受访者再次发送提醒以及问卷。事实证明，跟进是一个提高应答率的好办法。读者可以参考迪尔曼（Dillman，1978）中的提高信件调查中的应答率的其他办法。

抽样调查的应答率被定义为有效成员完成的问卷数量与样本中有效成员数量之比。这一定义虽然非常直观，但在遇到空白元素和外来元素的时候，我们还是会遇到一些问题。根据定义，此类非有效成员应当被排除在分子和分母之外，但判断一个成员是不在家还是空白元素却并不总是可能的。因此，举例来说，对于一个使用随机拨号调查时抽取到的一个号码，当重复拨号没有应答时，可能该号已被停用，为空白号，或者打电话时该住户不在家。类似地，在一个对18岁到24岁的年轻人的调查中，一个没有应答的住户可能包括也可能不包括一个以上我们目标群体的成员。在实际中，这种情形在不同的调查中的处理方式不同，从而导致了不可比的无应答率。因此，应当谨慎对待汇报的应答率，尤其看

看它们是如何计算的。

现在,对于并不复杂的由非政府调查机构实施的面访的应答率在70%到75%之间,并且其变动程度根据调查条件有所不同。一般来说,拒访是无应答的主要原因,然后就是不在家。较面访而言,电话访问常常有更低的应答率,其中拒访是主要原因。电话访问同时也有很多中断访问的情况(break-off),即在访问中间受访者停止了访谈。信件访问的应答率变动很大,从10%到超过90%。这一波动部分地依赖于跟进的程度,以及调查主题与受访者的相关性。

以目前的拒访率来看,不能忽略无应答偏误的风险。更重要的是,常常有证据表明无应答并非均匀分布在不同群体中,而是在一些群体中更加严重。比如,面访中的无应答率在城市内部比城市的其他地方更高。因为无应答率的不同,不同子群体中所抽取的样本的分布将与我们预期的不同。当这些群体特征与调查的变量相关时,以上偏差就会引起无应答率的偏误。如果我们能得到不同子群体的无应答率,那么我们仍然可以尝试通过使用加权调整对这一偏误进行调整,这一方法我们将在下一章中提到。然而,人们应当注意到,这些调整仅仅是对已知的不平衡的分布进行了调整,但不一定能够消除——甚至降低——无应答偏误。仅仅当无应答率为调查变量的总样本中各个子群体的一个随机子集时,这一方法才能消除无应答偏误,而这一情形在实践中基本是不可能的。因此,虽然使用加权调整可以让我们努力消除无应答偏误,但这并非解决此问题的完美方案。使用加权调整并不能削弱在数据搜集中追求高应答率的努力。

项目无应答,即数据搜集中出现的不合适的断裂,其出

现有一系列原因。受访者可能不知道问题的答案，或者因为某些问题比较敏感、尴尬或者他们认为其与调查主题不相关时而拒绝回答。在调查访问的压力下，访问员可能错误地跳过了问题或者没有记录下来问题的答案。即使当答案被记录到问卷上时，它可能也是无效的，因为其与其他问题的回答不一致。项目无应答的变异依赖于该项目的特性以及数据搜集的模式。简单的人口学的项目无应答率很低，然而对收入与支出的项目无应答率则有10%以上；非常敏感或者很难回答的问题可能有很高的项目无应答率。

一个解决项目无应答率的方法是将分析限制在有回应的那些项目内。在单变量分析中，总无应答率和项目无应答率常常可以通过这种方式来解决。因此，我们可以使用公式9.1中应答者的均值来估计总体均值，其中 W_m 现在被定义为没有给某一条目提供应答的各个元素的数量——完全没有回答或者没有回答该项目——与合格元素的总数之比。因此，总无应答偏差在此与项目无应答偏差是被同等对待的。

与总体无应答中使用加权调整对应，人们可以使用很多填补方法(imputation)来解决项目无应答率问题。这一方法通过给缺失应答填补值实现，在此过程中会将问卷中应答者对其他条目的回答作为辅助。一个方法是根据人们对其他相关条目的应答来将样本分为不同层级(classes)，然后将该层级的该问题的应答者的均值作为该层级无应答者的值。这一方法能够部分弥补层级之间不同条目的不同无应答率的问题，并且在估计总体均值方面，它与对总无应答使用基于相同层级的加权调整是等价的。

使用层级均值进行填补的劣势在于,它扭曲了该条目的分布,在该条目无应答之处使用均值替换,使得该层的均值处出现突起(spike),从而减少了分布的方差。一个变通的方法可以避免这一问题,即将该层该条目的某一值赋予无应答处。美国国家统计局使用过这种方法,也被叫做传统热卡法(traditional hot deck method)。首先人们分出层级,并给每一层级用该条目的单一值进行赋值,这一赋值可能基于上一次调查得到。然后,当前调查的记录可以按顺序进行。如果该条目是有应答的,这个值就会取代该层所被赋予的值。如果得到了无应答,那么其取值则保留为之前该层中所存储的值。这一方法的主要优势在于其具有计算的经济性,因为所有的填补都是根据数据中的单个来源进行的。然而,这一方法的缺陷在于,一个应答值可能被分配到了多个无应答处;当在一个层级内,具有缺失值的一条记录后面跟着一条或一条以上的缺失值时,这种情况就会出现。这种方法的另一个变体能够最小化应答值的多种用途,其方式是先通过将所有记录值排序分层级,然后将应答者和无应答者进行匹配;这一方式也不需要设置初始值(start-up values)。这一方法是美国统计局进行当前人口普查的辅助调查(March Income Supplement of the Current Population Survey)使用的一个复杂填补方法的基础(Welniak and Coder, 1980)。

另一个填补的方法使用了回归方程来预测缺失值,具体而言,使用应答者在问卷中对其他条目的回答作为预测变量,以便从应答者样本中得到回归系数。我们可以通过使用回归方程得到预测值作为填补,然而这么做的后果是该条目的方差将会偏低,正如层级均值的方法的缺陷一样。对这一

方法的一个修正是在回归预测过程中加入随机残差来避免对方差的低估。

填补的一个重要作用在于我们可以得到一个没有缺失值的数据集,而这能够极大地帮助我们的研究。然而,调查研究者需要意识到此过程中用到了填补。被填补的值应当在数据中被标出,从而让分析者能够区分真实值与填补值。与原始数据对比,一个包含了填补值的调查数据得到的结果应当得到严格的审查。其中的一个原因在于这一方法会带来估计量更高的、不合理的精确度。另一个原因则是虽然填补可能会降低单变量分析中无应答条目带来的偏误,但在多变量分析中,这一做法可能扭曲变量之间的关系,从而影响估计结果。对于一个填补方法的回顾以及这一方法对估计量的影响,请读者参考卡尔顿和卡斯普什克的研究(Kalton and Kasprzyk, 1982)。

第10章

调查分析

调查数据的分析可以使用很多种统计方法。这一部分并不回顾这些内容,而仅仅讨论与复杂抽样设计相关的分析方法。接下来要讲的分别是调查分析中权重的使用以及计算复杂抽样设计中估计量的抽样误差的方法。

第 1 节 | 权重

在抽样调查的分析中，人们一般用权重来给一些元素赋予相对其他元素而言更高的相对重要性。当抽样中使用了不等概率抽样时，我们就需要使用加权的方法；另外，这一方法在事后分层（poststratification）以及调整总体无应答时也会用到。下面，我们从介绍一个非等概率抽样设计中的加权方法开始，然后介绍其他的应用。

为了说明加权是如何实现的，我们首先考虑一个样本量为 10 的小样本。为了对一个大学的学生进行抽样，假设我们手中仅有的名单为每个课程的注册名单的总和。然后，我们从这些名单中抽取一个等概率的名单的样本——比如系统抽样——然后将被抽到的名单中的学生作为样本。假设名单的总数为 970，我们按照 1∶97 的比率抽取就可以得到一个包含 10 个名单的样本。但因为大多数学生选取了不止一门课，而学生在他们所选的课程中的编号是不同的，我们得到的名单的等概率抽样样本就会产生一个学生的非等概率抽样样本。某个学生选课的数量越多，其被选择的概率就越大。

假设抽样调查的一个目的是估计学生购买的教科书的平均数量，而表 10.1 给出了 10 个学生购买的教科书的总量

以及他们所上的课程数量，从而购买的教科书的简单均值为 $\sum \frac{y_i}{n}=\frac{47}{10}=4.70$。但这显然是一个对该大学所有学生购买教科书数量的均值的有偏估计，因为样本中学生被选择的概率是不等的。通过考察表 10.1 中的数据，我们看到学生选择的课程越多，其购买的教科书也会越多，从而简单的均值会高估总体均值。为了解决非等概率抽样样本的问题，我们使用与被选择概率成反比的权重来处理。如果一个元素的被选择概率为 p_i，那么其权重应当为 k/p_i，其中 k 为任何为了方便而选择的常数。

表 10.1　包含 10 个学生的样本中购买教科书的数量和选修课程的数量(假设数据)

学生号码	教科书的数量(y_i)	选修课程的数量(r_i)	权重 $w_i=12/r_i$	$u_i=w_iy_i$
1	2	1	12	24
2	5	2	6	30
3	6	3	4	24
4	8	3	4	32
5	3	2	6	18
6	7	4	3	21
7	6	4	3	18
8	3	2	6	18
9	5	3	4	20
10	2	2	6	12
	47		54	217

显然，我们可以选择 k 为 1，从而权重为 $1/p_i$。因此，当我们以 1∶97 的比率来在样本中抽取时，被抽取的学生的权重应当为 $97/r_i$，其中 r_i 为第 i 个学生选取的课程数量。因此，学生 1 的权重为 97，学生 2 为 48.5，学生 3 为 32.3，依此

类推。当估计总体的总数时，比如估计该校学生购买的教科书总量时，使用 $k=1$ 比较有用，因为 $k=1$ 时总体的总数可以通过加权总和 $\sum w_i y_i$ 计算出来。然而，被选概率常常非常小，并且不容易被处理，此时我们就可以选取其他 k 的值来简化权重。当我们采取 $k=1$ 以外的其他值时，加权样本总和 $\sum w_i y_i$ 就需要除以 k 来估计样本总数；然而，我们并不需要对均值、比率、方差以及其他对样本采取均值的统计量做其他调整。

另一个明显的选择权重的方法是将权重设为选课数量的倒数，$1/r_i$，因为这一变量是使得每个学生被选择概率不等的原因。对于第一个学生，这一权重为 1，第二个学生为0.5，第三个学生为 0.33，依此类推。这一方法潜在地预设 $k=1/97$，是完全可以接受的，但是这要求对 1/3 的取值四舍五入。为了避免这一点，在表 10.1 中的权重被设为 $12/r_i$（潜在预设 $k=12/97$）。当我们使用了这一权重时，样本均值则为：

$$\bar{y}_w=\sum w_i y_i/\sum w_i$$

这里，$\bar{y}_w=217/54=4.02$，这比我们之前估计的有偏的简单均值 $\bar{y}=4.70$ 小了很多。

因为 $w=\sum w_i$ 作为 $\bar{y}_w$ 的分母并非固定而是随着不同样本而变化的，所以加权均值是一个比率均值。正如第 6 章所讨论的，比率均值是总体均值的有偏估计，但是当分母的变化系数小于 0.1 时，这一偏差可以被忽略。将名单的样本作为简单随机抽样并且忽略 FPC 项，我们可以用以下方式估计权重的方差：

$$v(w)=ns_w^2=n\sum(w_i-\bar{w})^2/(n-1)=\frac{624}{9}=69.33$$

因此，w 的变异系数的估计值为：

$$cv(w)=se(w)/w=\sqrt{v(w)}/w=8.327/54=0.15$$

尽管该系数超过了 0.1，但它已经小到足够保证比例均值的偏差并非不可接受。因为这一系数会随着样本量的增大而减少，因此当我们有实际上会大很多的样本时，这一偏差不是大问题。

正如第 6 章所讲的，估计的加权均值的方差也是比例均值的方差。根据比例估计量的理论应用，$\bar{y}_w$ 可以被写为 $\sum u_i/\sum w_i=u/w$，其中变量 u_i 被定义为 $u_i=w_iy_i$。然后，给定 w 的变异系数小于 0.2，$\bar{y}_w$ 的一个近似的方差的估计则可以为：

$$v(\bar{y}_w)=[v(u)+\bar{y}_w{}^2v(w)-2\bar{y}_wc(u,\ w)]/w^2$$

而上式仅仅是将我们现在的表示法代入了公式 6.5。使用表 10.1 的数据，我们可以做下列计算：

$$v(u)=n\sum(u_i-\bar{u})^2/(n-1)=3\,241/9=360.11$$

$$c(u,\ w)=n\sum(u_i-\bar{u})(w_i-\bar{w})/(n-1)=52/9=5.78$$

因此

$$v(\bar{y}_w)=\frac{\left[\frac{3\,241}{9}+\left(\frac{217}{54}\right)^2\frac{624}{9}-2\left(\frac{217}{54}\right)\left(\frac{52}{9}\right)\right]}{54^2}=0.491\,5$$

并且 $se(\bar{y}_w)=0.70$。

有必要比较这一非等概率抽样样本的精确度与具有相同样本量的简单随机抽样样本的精确度。为此,我们需要一个购买教科书数量的方差的估计量。我们可以根据公式10.1给出:

$$s_w^2 = \frac{n}{n-1} \frac{\sum w_i (y_i - \bar{y}_w)^2}{\sum w_i} = 4.382 \qquad [10.1]$$

因此,对于一个包含10个样本量的简单随机抽样,忽略FPC项,通过 $v(\bar{y}_0) = \frac{s_w^2}{10} = 0.438\,2$,可以估计出样本均值的方差。因此非等概率抽样样本的估计设计效应为:

$$d^2(\bar{y}_w) = \frac{v(\bar{y}_w)}{v(\bar{y}_0)} = \frac{0.491\,5}{0.438\,2} = 1.12$$

说明不等概率抽样使得方差增加了12%。在抽样框不完美的情况下采用不等概率抽样所带来的精度损失是正常的,并且当抽样概率变化很大时,这一损失可能非常可观。因此,当我们面临这种情况时,应当尽量避免选择概率的变化太大。

作为在抽样框不完美情况下需要使用加权的第二个例子,我们考虑元素的群构成的抽样名单。假设在某个城市中我们从 A 个住址中抽取了一个住址的等概率抽样样本,然后我们使用基什选择表在每个被选择住址中随机抽取一个成人。因此在住址 α 抽取到成人 β 的概率如下:

$$P(\alpha\beta) = P(\alpha)P(\beta \mid \alpha) = (a/A)(1/B_\alpha)$$

其中 B_α 是在住址 α 居住的成人数量。为了解决选择成人的概率不等的问题,我们需要在分析中使用与 $1/P(\alpha\beta) =$

AB_{α}/a 成比例的权重。一个明显的加权方法是将每个被抽到的成人的权重设置为其住址包含的成人数(比如 B_{α})(尽管这些权重在理论上是必须的,但在实际中人们通常很少用到,因为它们非常小,每个住址包含的成年人个数的变异性并不大,因此它们通常仅仅对调查统计量有微乎其微的影响。参见 Kish, 1965:400)。

另一个会导致不等概率抽样的抽样设计是非比例分层。第 4 章已经讨论了一个总体均值的估计可以通过先在每一层计算样本均值,然后将这些估计通过加权平均 $\bar{y}_{st} = \sum w_h \bar{y}_h$ 合并起来。一个替代性的方法是为每一个被抽取的元素分配权重,在每一层中给所有的元素相等的权重,但是不同层之间的权重不同,然后我们可以使用 $\bar{y}_w$。这些权重与每一层中被选择概率的倒数是等比例的。比如,$w_{hi} = kN_h/n_h$ 为层 h 中被抽取的元素的权重。因此我们有:

$$\bar{y}_w = \sum_h \sum_i w_{hi}\, y_{hi} / \sum \sum w_{hi}$$
$$= \sum_h k\, N_h \bar{y}_h / \sum_h k\, N_h = \sum_h w_h \bar{y}_h$$

所以 $\bar{y}_w$ 和 $\bar{y}_{st}$ 是相等的。在此情况下,$\bar{y}_w$ 并非比例均值,因为其分母为常数。使用 $\bar{y}_w$ 而不是 $\bar{y}_{st}$ 有一个计算上的便利:一旦我们设置了权重,我们就可以使用加权数据的标准电脑程序来得到调查统计量。

加权也可以在选择后分层(stratification after selection)或事后分层(poststratification)的方法中实现。通过这一方法,一些补充变量的总体分布的信息会被用到增加样本估计量的精度的分析中。因此,比如,当我们从最近的人口普查中知道了总体中年龄的分布,样本就可以按照年龄组被划

分，在每个年龄组中计算调查变量 y 的均值（$\bar{y}_h$），这些均值联合起来可以得到总的估计量 $\bar{y}_{ps}=\sum w_h\bar{y}_h$，其中 w_h 为总体中年龄组 h 的比重。在不等比例分层中，事后分层的均值也可以通过加权均值的方式表示出来，其中每个元素具有与 N_h/n_h 成比例的权重。如果忽略无应答与无覆盖（noncoverage）的问题，事后分层能够调整层之间可能由于几率不同而不同的样本分布，从而使得其服从一个已知的总体分布。给定一个事后层（poststrata）的期望样本量为 10 个或 10 个以上，事后分层的均值的方差大致等于基于相同层的等比例分层的均值的方差。在 w_h 为已知而每个元素属于哪一个层在选择阶段不能被确定的情况下，事后分层是有用的。在这种情况下，事前分层（prior stratification）不能被使用，但人们可以通过被抽取的元素来搜集信息，使其能够被分配到某一层中，从而可以使用事后分层。事后分层也可以被充分应用于抽样设计阶段使用的分层因子之外的因子中。对于按比例分层而言，当调查变量的层间有一定异质性，也就是说层中有同质性时，事后分层所获得的精确性可以累积。

将样本进行加权调整到一个已知的总体分布不仅仅会对样本波动产生影响，也会对无应答以及无覆盖（一些元素没有被包含到抽样框中）产生影响。比如，当无应答率在年轻人中间更高时，或者当他们更多地在抽样框中缺失时，对样本进行加权从而保证其服从一个已知的年龄分布可以解决这些问题。然而，这里人们应当注意到，这一方案是通过将应答者在给定的年龄组内进行加权实现的。对于每个年龄组之内的应答者和无应答者在调查变量方面的区别而言，一些无应答的偏差仍然存在。

与事后分层类似，对无应答或无覆盖的事前加权调整要求人们从外部来源获得一些辅助变量的总体分布的信息，比如年龄。无应答调整的另一个类型仅仅依赖于样本中的信息，但这一信息必须是对应答者和非应答者都适用的。元素所在的层的信息或者初级抽样单位的信息常常被用做这一类型的调整。比如，假设样本被按照地理区域划分，而在某一区域之内，将其按照该元素是否属于农村、郊区或者城市中心的位置划分。给定一个等概率抽样的样本，对无应答率的变动在组间的调整可以通过给层 h 的应答者分配 n_h/r_h 来实现，其中 n_h 为被选取的总样本量，而r_h 为该层被访者样本量。这些调整使得受访者样本的分布服从总样本的分布，其中层中的应答者被加权来代表该层的无应答者。这种类型的调整仅在无应答时进行，而非无覆盖。

在实际中，设置权重可能是一项非常复杂的任务，因为人们往往需要做一系列的调整。但首先，可以设置权重来调整不等抽样概率，然后将其根据样本中某一层中不同的应答率来进行进一步调整，最后再采取一些修正使样本分布使其服从一个已知的总体分布。在设置权重时应当非常小心，因为在此过程中很容易出现严重问题。

第 2 节 | 抽样误差

正如我们已经在不同的抽样设计中讨论的，抽样估计量的抽样误差（sampling errors）程度依赖于调查的抽样设计。统计学以及大多数电脑程序中的标准误差公式仅仅与无限制抽样（unrestricted sampling）有关（有放回的简单随机抽样）。不应当无条件地将这些公式应用到其他抽样设计中，否则可能导致高估，或者更常见的是会低估抽样误差。

在无放回的简单随机抽样情形下，样本均值的方差比具有相同样本量的无限制样本的均值的方差小，并且其比率为 $(1-f)$，即有限总体修正项（finite population correction term）。当总体较大时，抽样比率 f 常常比较小，FPC 近似于 1。在这种情况下，可以在简单随机抽样设计中放心地使用无限制抽样的标准误差公式。

一个在层内使用简单随机抽样的比例分层抽样设计给出的统计量，至少与使用简单随机抽样设计得到的统计量具有一样的精确度。由于层内元素具有相对抽样变量而言更高的同质性，这些估计量的精确度实际上更高。这种设计的无限制抽样的标准误差公式因此会趋向于高估统计量的抽样误差。当 FPC 项能够被忽略以及通过分层所获得的精确度很微小时，仅使用无限制抽样的标准误差公式即可。然

而,在完全信赖这些公式之前,最好检查一下忽略分层得到的精度收益是否合理。

有两方面使得非比例分层抽样的情况更加复杂。第一,由于非比例分层是非等概率抽样的,使用无限制抽样的标准误差公式需要使用总体参数的加权后的估计量。比如,无限制样本中的样本均值的标准误为 $\sigma/\sqrt{n}$;在非等概率抽样情况下使用这一公式,总体方差 σ^2 应当通过公式 10.1 中的加权项 s_w^2 估计出来。第二,非比例分层对调查估计量精度的效果并不像它在等比例分层中一样一目了然:非比例分层抽样得出的估计量与具有相同样本量的无限制样本得出的估计量相比,可能更精确或者可能更不精确,具体情况则取决于样本在不同层之间的分配。假设调查元素的成本对于所有层都是一样的,对于估计某一变量的总体均值的样本最优的分配会产生至少与其等比例分配情况下相同的精度,而当该变量在不同层之间的元素方差发生变化时,其精度还会更高。这一无限制抽样的标准误差公式因此会趋向于高估均值的抽样误差。然而,使用无限制抽样的标准误差公式会趋向于低估这一设计下的其他估计量的抽样误差。

非比例分层常常被用来分别提供不同领域研究的估计量,其中层代表了以更高比率抽取的小的领域,从而使其在领域内具有足够的样本。这种应用方法常常导致总体估计量的精度损失,并且当以更高的比率在一些领域抽样时,这一损失更严重。举一个简单的例子,考虑两个层,每层都是一个单独的研究领域并且人们对其分别估计估计量,其中一个层包含 90%而另一个层包含 10%的元素,并且为简单起见,假设这两个层具有相同的均值和方差。如果从每个层中

抽取具有相同样本量的样本，我们就需要两层比率为 9∶1 的权重来得到总体的估计量。如果忽略 FPC 项的话，与具有相同样本量的非限制样本相比，总体样本均值的方差就被提高了 1.64 倍。当需要通过显著变化的不同权重来调整不等概率抽样时，就可能会导致精度的可观损失。其结果是，使用非限制抽样的标准误差公式可能会严重低估抽样估计量的误差。

当群的层内相关性系数 ρ 为正时，与具有相同样本量的简单随机抽样相比，群就会有精度损失。这一损失既依赖于 ρ 的大小也依赖于每个群被选择的平均子样本大小，正如我们在第 5 章中所讨论过的。当平均子样本量比较大时，即使 ρ 比较小，这一损失也会相当严重。非限制抽样的标准误差公式会趋向于低估多阶段群样本所得到的估计量的抽样误差，而且常常是可观的低估。

在实践中，抽样设计往往是很复杂的，既包含了多阶段抽样，也包含了在每一抽样阶段的某些形式的分层。常用的设计包括按比例抽样以及等概率抽样，或者近似等概率抽样。通常来说，这些统计量的抽样误差计算的实证结果，是由于群导致的精度损失会大于由于按比例分层所得到的收益，从而使得复杂抽样设计与具有相同样本量的无限制样本相比，会产生较低精度的估计量；即，设计效应大于 1。设计效应的大小依赖于一系列因素，包括群的特征、每个群的平均子样本量、所使用的分层方式、研究的变量以及统计量的形式。因此，在全国范围内的概率样本，一些基本人口变量，比如年龄和性别，其均值和比重的设计效应一般接近于 1，表明了地理群对于这些变量表现出非常小的内部同质性。然

而，对于社会经济变量以及相关变量的设计效应一般会大于1，因为社会经济地位相同的人更可能居住在一起。对于总体的子群均值或比例来说，设计效应在群之间基本是平均分布的，或者说是跨群的，它们基本小于基于总样本的均值或者比例的设计效应。两个子群均值的设计效应的差别一般小于子群的均值的设计效应本身。回归系数的设计效应一般与均值间设计效应的差相似。然而，不管对于什么统计量，复杂抽样设计的设计效应基本总是大于 1 的，这一效应有时很小，但有时很大。因此，使用非限制抽样的标准误差公式一般会高估抽样结果的精度。

近年来，人们发展了一系列电脑程序来计算复杂抽样设计下统计量的抽样误差。从卡普兰和弗朗西斯的研究中(Kaplan and Francis，1979)，我们可以看到这种程序的一个列表。在大多数情况下，这些程序将初级抽样单位(PSU)作为有放回的抽取，尽管在实践中人们往往采用无放回抽取。将初级抽样单位当做有放回的抽取会高估方差，但是这一高估是很微小的，因为我们考虑到第一阶段抽样比率很小。有放回抽样假设的重要优势在于计算的经济性以及假设的慷慨性。正如在第 5 章中所说的，如果第一阶段的抽样比率很小，样本均值的标准误可以简单地通过初级抽样单位总体的变化估计出来；人们并不需要在初级抽样单位内部根据抽样的变化来估计，而这一点节省了人们很多计算上的工作。然而，更重要的在于，这一假设非常慷慨：在有放回抽样的假设下，可以应用对于特定估计量的标准误差模型，不管初级抽样单位内部采取了什么子样本的方式。因此，当元素按照(1)在被选的初级抽样单位内使用简单随机抽样；(2)使用系

统或分层抽样或(3)采用进一步的抽样或分层阶段抽取时，相同的公式都可以被使用。这一慷慨性非常有吸引力，不仅在于一个简单的程序就能产生出任何形式的子样本设计的估计量的标准误，更在于该程序的用户并不被要求根据抽样设计来应用这一程序。这些程序的使用仅仅需要每个抽样数据的记录包含其属于哪一个初级抽样单位的指令，以及第一阶段分层的信息。

计算复杂抽样设计的估计量的抽样误差有一些一般性的方法。其中之一就是泰勒展开(Taylor expansion)或者德尔塔方法(delta method)，正如我们在第 6 章中讨论比例均值时提到过的(见公式 6.5)。这一基本方法是获得一个对估计量的线性近似来获得估计量的方差。对于一些简单的估计量而言，这一方法使用起来非常方便。很多复杂样本设计中的计算样本均值、比例、子群均值和比例以及均值和比例的差的方差均使用了这一方法。正如在第 6 章中提到过的，对于比例均值或者比重的方差的估计对泰勒展开法的合理使用要求比例的分母的变异系数小于 0.2。大多数程序在它们的输出中提供这一系数的值；特别是对于子群的分析而言，常常需要检查这些值以保证它们足够地小。

估计标准误的另一个方法是将样本设计为允许标准误估计在任何估计量下都能被计算。正如在第 7 章中提到的，重复抽样方法能够通过将总样本构造为一系列独立重复的，每一个都是相同样本设计的联合体来实现这一目标。每个重复估计的变动因此可以作为联合样本的标准误的估计基础，不管估计量或者重复样本的设计如何复杂。正如之前讨论过的，使用具有多阶段设计的简单重复抽样的重要缺陷在

于，为了估计标准误达到一定精度所需要的足够的重复抽样次数，以及人们希望得到很多分层来获得精确的调查估计量之间的矛盾。有鉴于此，人们很少使用简单重复抽样。作为一种替代，人们发展了伪重复抽样（pseudoreplication）的技术，其采用了简单重复标准误估计量的优势，能够尽可能地提供对标准误的估计，同时避免了对分层的限制。我们会简单介绍平衡重复复制（Balanced Repeated Replications，BRR），其有时也会被称为半样本复制（half-sample replication）（Kish and Frankel，1970，1974；Frankel，1971；McCarthy，1966）。

BRR 方法常常与成对选择设计（paired selection design）一起使用，在每个层中选取两个初级抽样单位。正如之前说过的，在很多多阶段设计中，将初级抽样单位们分层直到每个层选择出一个初级抽样单位，此时我们就需要用折叠层法来估计方差；成对的折叠层近似于配对选择设计的方法。在 BRR 方法下，从每个层中选择的两个初级抽样单位会被当做独立选取的一样。从重复抽样的角度看，这一样本可以被当做由两个重复的样本组成，其中一个包含两个初级抽样单位中的一个，即从每个层中随机选择的，另一个则包含剩余的初级抽样单位。如果 z' 表示基于第一个复制样本，或半样本的参数 Z 的样本估计（比如，回归系数），而 z'' 表示基于另一半样本，或补充样本的对应估计，那么根据重复抽样理论，我们就可以根据公式 7.1 计算 $\bar{z}=(z'+z'')/2$ 的方差，其中 $c=2$

$$v(\bar{z})=[(z'-\bar{z})^2+(z''-\bar{z})^2]/2 \qquad [10.2]$$

在实际中，总体上用来估计 Z 的是 $\tilde{z}$，方法是将两个一半的样本合并起来，但是 $\bar{z}$ 和 $\tilde{z}$ 通常极为接近。因此，作为

一个近似，在上述方差的估计式中，$\bar{z}$ 可以用 $\tilde{z}$ 替代。

简单重复复制方差的估计量的局限性在于，它仅仅基于一个自由度，在实践应用中的稳定性不够。BRR 中解决这一问题的方案是重复从母样本中构造半样本的过程，每次都计算出方差估计量，然后计算得到方差估计量的均值。因此，如果 z'_t 表示基于第 t 个半样本的 Z 的估计值，而 z''_t 为基于其对应的另外半个样本的估计值，则 $\tilde{z}$ 的方差估计量可以由下式给出：

$$v(\tilde{z}) = \sum[(z'_t - \tilde{z})^2 + (z''_t - \tilde{z})^2]/2T$$

以上平均是基于 T 个半样本以及其另一半得到的，并在公式 10.2 中使用 $\tilde{z}$ 代替 $\bar{z}$。

以上部分解释了 BRR 的“重复复制”部分。其中，“平衡”部分指的是半样本的选取方式。T 个半样本并不是被独立抽取的，而是以平衡的方式被抽取的，从而产生一个总体方差的有效估计量。为了实现总体平衡，被选择的半样本数量 T 需要大于等于层的数量且需要是 4 的倍数。因此，比如，当我们有 22 个层时(比如，在成对选择设计中有 44 个 PSU)，$T=24$ 个半样本就能够实现总的平衡；当我们有 47 个层时，就需要 $T=48$ 个半样本。如果 z_t 的计算需要大量的工作，并且当层的数目相当多时，为了实现总体平衡所需要的对所有半样本的计算就可能是惊人的；在这种情况下，我们就可以在不同的技术下少使用一些半样本来实现部分平衡。

刀切重复抽样(Jackknife Repeated Replications，JRR)是在复杂抽样设计下的另一个方差估计方法(Frankel，1971；Kish and Frankel，1974)。像 BRR 一样，它使用了重复复制

方法。在 JRR 方法下，人们通过扔掉单个初级抽样单位，并且将该层中的其他初级抽样单位加权，从而实现复制来保留层之间的样本分布。这一操作会被重复数次，每次扔掉一个不同的 PSU。当被抽取的初级抽样单位总数 $\underset{\sim}{a}$ 非常小的时候，其中的每一个就可以被依次扔掉来产生 $\underset{\sim}{a}$ 个复制，但这一工作并非必须要全部完成。唯一的要求在于，每一层需要有至少一个初级抽样单位被扔掉；如果对于一个或多个层，这一条件没有被满足，这些层所贡献的方差就不会在总体方差估计中被代表。令 z_{ht} 表示基于根据层 h 生成的第 t 个复制得到的 Z 的估计，对于 $\tilde{z}$ 的一个 JRR 方差估计量则由下式给出：

$$v(\tilde{z}) = \sum_{h=1}^{H} \sum_{t=1}^{t_h} (a_h - 1)(z_{ht} - \tilde{z})^2 / t_h$$

其中 a_h 为层 h 中所抽取的初级抽样单位数目，t_h 为从层 h 中扔掉初级抽样单位得到的复制的数量。当样本中的每个初级抽样单位都被依次扔掉后，$t_h = a_h$。 正如这一公式表明的，JRR 优于 BRR 的一个地方在于，它能够很容易地应用于除了成对抽样设计（即 $a_h = 2$ 时）以外的其他设计中。

所有之前的方差估计方法都包含了近似，但是模拟研究表明它们都能得到令人满意的结果。对它们的选择很大程度上取决于计算的成本、程序的可得性、待估计量的适用性以及相应的抽样设计。泰勒展开方法常常对于简单估计量更普遍，BRR 和 JRR 方法的优势在于其适用于复杂的估计量。BRR 方法的应用主要限于配对选取设计中，但这一设计在实践中也适合于绝大多数样本。当不适合的时候，就可以使用 JRR。

抽样调查通常针对非常多的变量进行数据搜集，并且会产生无数变量结果及其之间的关系。即使我们有抽样误差的电脑程序，计算一个调查报告中的所有估计量的标准误也是不可能的；即使有可能计算，最终的报告也会因此而过于冗长。由于这些原因，调查分析常常仅计算主要结果的标准误，并选择性地报告其他的标准误。之后这些计算可以用于发展更加一般性的模型，在那个时候再推断其他的标准误（参见 Kish，1965：574—582）。

抽样误差的实际估计的更多细节以及一般性抽样误差模型的使用可以参考卡尔顿的研究（Kalton，1977）。基什和弗兰克尔（Kish and Frankel，1974）的论文讨论了调查样本的抽样误差估计方法并提供了比较泰勒展开、BRR 和 JRR 方法模拟的结果。

第11章

样本量

进行抽样设计首先要面对的一个问题是,“我们需要多大的样本量”? 这一问题的讨论之所以现在才开始进行,是因为它与很多之前介绍的内容有很大联系。

为了描述基本的想法,我们举一个面对面访问的简单例子,其目标是估计一个具有 15 000 成年人的城市在新图书馆落成后回答将会使用新图书馆的人所占的比例。为了决定一个合适的样本量,首先需要设定我们对估计量所需要的精度。这并不容易,因为人们开始所设定的精度要求一般都会被高估。比如,假设最初的设定要求一个估计量在 2%的总体人口中有 95%的可能性;换句话说,95%的置信区间应当为样本百分比加减 2%。这一设定要求 1.96SE(p) =2%,其中 p 为样本百分比。假设开始时我们使用简单随机抽样,并忽略 FPC 项, $SE(p) \approx \sqrt{PQ/n'}$,其中 P 为总体百分比,$Q = 100 - P$,n' 为最初估计的样本量。因此 $1.96\sqrt{PQ/n'} = 2$ 或者 $n' = 1.96^2 PQ/2^2$。为了决定 n',我们需决定 P 的值。由于 PQ 在 $P = Q = 50\%$ 的时候最大,一个保守的选择是将 P 设定为尽可能接近 50%。假设我们认为 P 可能在 15%到 35%之间,那么保守的选择为 $P = 35\%$。有了这个选择,我们得到 $n' = 2\,185$。如果这一原始样本量与总体相比很小,可

以忽略 FPC 项，这一样本量就是可以的。然而，在目前的例子中，不能忽略 FPC 项。一个修正的样本量的估计方法是将 FPC 项考虑在内，其中 $N=15\,000$，如下：

$$n=Nn'/(N+n')=1\,907$$

上述的计算假设了简单随机抽样，我们需要对其他样本设计进行修正。这一修正包含了将简单随机抽样样本量与其他复杂设计中的抽样效应的相乘项。如果该城市的成年人列表在上述例子中是可能的，那么我们可以使用一个非聚类的等比例分层样本（unclustered proportionate stratified sample）。在这种情况下，通过分层我们得到了更多的精度，一个小样本就够了。然而，正如之前说过的，估计百分比时的等比例分层的收益一般很小，从而样本量的缩减仅仅是细微的。在目前的例子中，比如等比例分层设计下的样本百分比的设计效应的估计值为 0.97，那么对于非聚类的等比例分层设计的样本量会给出在 ±2% 置信区间内为 $0.97\times1\,907=1\,850$。

如果我们没有该市成年人或其住址的列表，我们可能需要进行地区抽样（area sampling），首先抽取城区，然后列举城区中的住址，从中抽取住址，最后从被抽取的住址中抽取一个或多个成年人。在这一抽样设计中，基本上必然会用到分层以及 PPS 选取。假设一个分层多阶段样本，我们从每个初级抽样单位（街区）中平均抽取 10 个人，那么预期的设计效应就是 1.3。因此这一设计要求的样本量应当是 $1.3\times1\,907=2\,479$。

另一个在计算中需要被考虑的因素是无应答。假设预

测的应答率为 75%,那么为了实现多阶段抽样的样本量是 2 479 个成年人,所需要的样本量应为 2 479/0.75=3 305。当然,这一调整仅仅是为了得到理想样本量,它并不能解决无应答偏差的问题。

在这一点上,研究者可以回顾最初的精度设定来看能否将要求放宽。假设回顾时研究者将置信区间放松为±3%,样本量从而可以被减少为 1 581。在实践中,对某个估计量的精度要求并不是固定的。因此,样本量也常常基于调查成本与精度水平的粗略评估。但应当注意到,所选的样本量取决于一些预估的量,比如汇报将使用图书馆的人的百分比、设计效应以及无应答率。对这些量预估产生的误差会导致抽样估计量与我们设定的精度有偏差,但这仅仅是反向作用,估计量对于总体参数而言仍然是可信的。

在固定了样本量之后,下一步是决定抽样比率。如果我们从一个有 15 000 成年人的城市抽取样本,就需要考虑名单上有空白元素(死亡或者搬离了该地区)和外来元素的可能,以及处理缺失元素时使用链接程序的后果。如果说名单中的 4%为空白元素并且没有使用任何链接,抽样比率就应当为 2 479/(0.96 × 15 000) = 0.172,或者 1/5.81,从而得到 2 479 的样本量。在实践中,可以为了方便将这一抽样比率四舍五入,比如 1/5.8 甚至 1/6,从而得到相应的预期样本量分别为 2 483 或者 2 400。

在多阶段地区样本中,抽样设计要求一个住址的样本,其中每个住址抽取一名成年人。假设在最近的人口普查中该城市包括 6 500 个住址。首先应当更新这一数字来修正从人口普查日到调查日之间的变化,同时修正两次普查之间任

何城市边界的变化。假设由于这些变化，目前住址的数量为 6 750。除此以外，人们也需要考虑由于调查的抽样过程可能不能像人口普查一样完整地覆盖该城市；比如，样本的覆盖率可能为人口普查的 95%。如果使用这一数字，为了得到所需样本量为 2 479 的样本，我们就需要抽样比率为 $2\,479/(0.95\times 6\,750)=0.386\,6$，或者 1/2.59 个住户。像以前一样，抽样比率可以四舍五入为 1/2.6，从而得到预期样本量为 2 466。

虽然上述例子表明了选取样本量会遇到的几个问题，但它还是过于简单了。在实际调查中，抽样调查常常是多目的的，需要考虑很多估计量。另外，不仅在总样本中需要这些估计量，一些子群体也需要，比如一个国家的不同区域，不同年龄层的群体或者不同教育程度的人们，等等。很多研究中需要大样本的主要原因是，需要针对子群体的估计量保证足够的精度或者比较不同子群体之间的估计量。更大的样本能够使子群体样本的分析更精确，并且样本量越大，分析就越细致。样本量的选择常常取决于增大样本量导致的成本的增加以及分析上的收益的取舍。

第12章

两个例子

这一章提供了两个例子来说明在实际应用中如何综合前面的技术。一个例子是对美国的全国性面访的抽样设计，另一个是对电话访问的抽样设计。

第 1 节 | 全国性面访调查

密歇根大学的调查研究中心(Survey Research Center, SRC)以及芝加哥大学的国情调查中心(National Opinion Research Center, NORC)都对个人、家庭、住户以及其他单位的面访有全国性的概率样本设计。它们每十年就会利用十年一度的人口和住房普查(Census of Population and Housing)的最新数据来修正这些样本的设计。在 1980 年普查之后,两个机构就一起合作,根据一个共同的抽样设计来选取它们的主样本。下面我们会描述这个设计,这是一个分层多阶段地区样本,其中在不同阶段使用了按估计容量比例的概率抽样(PPES)。

NORC/SRC 全国性抽样设计的初级抽样单位(PSU)是标准都市统计区(Standard Metropolitan Statistical Areas, SMSA)、区县或者在 1980 年人口普查中人数最少为 4 000 的县的小组。样本中绝对包含了 16 个最大的 SMSA(比如,纽约、芝加哥、旧金山、波士顿、圣路易斯以及亚特兰大)。它们是自代表(self-representing)的初级抽样单位,可以被当做层。根据以下程序,通过 PPES 从剩下的初级抽样单位中抽取 68 个初级抽样单位,其中规模的度量为初级抽样单位中在 1980 年人口普查的有人的住房数量。首先,将初级抽样

单位分为 68 个具有大约相等规模的层(比如,包含近似相等的居住单位的数量)。这些层首先通过将初级抽样单位分为四个普查区域(中北部、东北部、南部和西部),然后在区域中分为 SMSA 和非 SMSA 来实现。然后,SMSA 进一步按照地理位置以及其中最大城市的规模分层。非 SMSA 被按照地理位置和其总的规模分层。然后,从 68 个层中使用 PPS 抽取一个初级抽样单位;为了保证初级抽样单位在其他控制变量上的代表性,比如南部农村黑人的比例以及西部西班牙人的比例,我们对层间的选择方式进行了控制(对于控制选择技术的描述,参见 Goodman and Kish, 1950; Hess et al., 1975)。

接下来的抽样程序是在 16 个自代表的 SMSA 和 68 个被抽取的初级抽样单位中抽取小群。这里抽取的群是城市地区的街区,其中普查已经提供了街区的统计量以及其他地区的编号。这些群最少包括 48 个居住单位。对于小于这一最低限度的居住单位,应根据地理相邻原则将其合并。这些群包含了被抽取的初级抽样单位中的第二阶段单位,并且其中的六个从调查机构(SRC 和 NORC)的主要样本中每个被抽取的初级抽样单位中得到。在每个自代表的 SMSA 中,这些群实际上是初级抽样单位。对每个机构,我们在八个最小的自代表的 SMSA 中选六个群,但较大的自代表的 SMSA 则需要更多的群被抽取。比如,对于每个机构的主样本,在纽约 SMSA 我们抽取 24 个群,而在洛杉矶 SMSA 则抽取 18 个群。

在自代表的 SMSA 以及被抽取的初级抽样单位中选取群是通过 PPES 进行的,其中使用 1980 年有人的住房数量作

为度量规模。对于群的一个已排序的列表，我们使用系统抽样来得到对于排序时被使用的变量的模糊分层（implicit stratification）的收益。在自代表 SMSA 或者被抽取的初级抽样单位中，群被按照区县排序，具体通过次级划分（minor civil division）、普查地段（census tract）或者地区编码以及街区号码排列。区县则按照规模和地理位置被排序。对于具有规模信息以及次级划分（当地政府的单位比如城市和城镇）的家庭收入中位数的 20 个州，这一划分是按照规模和收入中位数排序的。街区和编码地区是按照普查地段号码排序的，然后按照街区或者编码地区号码来产生地理顺序。

根据这一程序产生的被抽取的群的规模变动很大，小到 50 个有人的居住单位，大到 700 个甚至更多。此后我们进一步在较大的群中进行抽样，从而将其规模缩小到一个可控的范围。这一步骤首先要求这些群被划分为定义清晰的部分，然后对每个部分分配其规模的约数。这一过程基于 SRC 和 NORC 调查人员对群的筛查，其中他们对有人的居住单位分布进行计数，然后从每个大群中使用 PPES 抽取一小部分。

抽样设计的最后一步是调查人员在所有被选择的部分中对居住单位进行列表。这些列表可以被用做很多调研的抽样框。从名单中抽样的方法可能在调研之间有很大差别，并且即使在大的部分和小的部分之间也会有差别。因此，根据时间有效性来更新名单是非常重要的。

对于其他关于全国性抽样设计的介绍，请读者参考美国人口调查局（U.S. Bureau of the Census 1978）或者基什（1965：第 9、第 10 章）关于 CPS 的介绍。

第 2 节 | 电话访问调查的例子

近年来，在美国人口调查中使用电话访问的情况越来越多，一部分原因是电话的高拥有率。在 93%的家庭拥有电话的情况下，电话号码成为很多调查的一个有效的抽样框。然而，在考虑电话访问的时候，人们应当注意到剩下的没有电话的 7%的家庭，因为这些家庭有很大一部分是低收入的，户主并非白人并且在 35 岁以下，或者居住在南方（Thornberry and Massey, 1978）。在研究者需要对这些群体也有足够代表时，电话访问样本可以在一个双抽样框设计（dual frame design）中与其他样本合并起来——可以是一个地区样本（参见 Groves and Lepkowski, 1982）。

当我们对一个既定的总体进行电话访问时，就面临了一个应当采用哪一个抽样框来抽取家庭的问题。一个明显的选择是已出版的电话号码簿，但是因为很多号码并不在其中，它并不充足。超过 20%的住户号码不在电话号码簿中，因为(1)他们是新近搬入的；(2)订阅者缴费使其电话不被公开或者(3)在号码簿的准备过程中出现了错误。考虑到这些遗失的元素所导致的潜在偏差，我们可以使用多种修正方式，比如先抽取电话号码，然后在其最后一位加上一个常数或者使用随机数代替号码的后两位（参见 Frankel and

Frankel, 1977)。然而,这些方法并不能够给每个家庭一个已知并且非零的被选择概率,而这正是概率抽样所要求的,因此,这可能引起估计偏差。

一个替代性的抽样框是采用所有可能的电话号码集。在美国,电话号码是由 10 位数字组成的,分为三个部分,比如 301—555—1212,其中第一部分是地区编码,第二部分是交换中心代码,而后四位则是后缀。总共有 100 多个地区编码以及 30 000 多个交换中心代码(比如,地区和交换中心代码的合并)在用。对于一个交换中心代码而言,有 10 000 个后缀可以用,但是其中的大多数是空号或者为商用的非居民号码。

基于这种抽样框的一种抽取居民电话号码的方式是,从一个地区码/交换中心码的组合中随机选取一个(一个随时更新这些组合的名单可以从美国电讯的长线部门得到),然后在 0000 到 9999 的范围内选取一个四位随机数来作为电话号码的后缀。随机数位拨号(Random-Digit Dialing, RDD)的一个简单版本会覆盖所有居民号码,但是它会有很多的空白元素(空号)以及外来元素(非居民号)。这些空白和外来元素当然可以简单剔除,剩下的样本从而构成一个居民号码的概率样本,但是由于为了剔除这些号码需要拨打很多电话,这一过程的费用高昂:平均而言,人们需要抽取五个号码才能得到一个居民号码。

另一个 RDD 方法可以减少没用的电话号码,其描述见瓦克斯伯格(Maksberg, 1978)。这一方法将电话号码的抽样框视做一个 100 个数字的库的集合,其中的库由地区码/交换中心码的组合以及后缀的前两位数定义。因此,每个地

区码/交换中心码组合之内,包括了 100 个包含 100 个数字的库,即后缀 0000—0099,0100—0199,0200—0299,…,9900—9999。这些库可以使用等概率抽样抽取,并且在每个库中随机抽取一个数字。如果该数字并非居民号码,该库就被拒绝;如果是居民号码,那么就可以安排对其的访问,并且可以在该库中继续抽取随机数直到特定数目的家庭被抽取到。

运用瓦克斯伯格方法(Waksberg scheme),选择并接受第 α^{th} 个库的概率与其包含的居民号码的比例成比例,即 $B_\alpha/100$,其中 B_α 为该库中居民号码的数量。给定库 α 被接受,一个特定的居民号码被抽取的概率为 $(b+1)/B_\alpha$,其中 b 为当第一个号码为居民号码时抽到的其他居民号码的数量。因此在库 α 中居民号码 β 的选择方程是:

$$P(\alpha\beta) \propto \frac{B_\alpha}{100} \times \frac{b+1}{B_\alpha} = \frac{b+1}{100}$$

因此,如果每个库中正好选取了 b 个额外居民号码,这一方法就是等概率抽样的。实际上,这些库是用 PPS 抽取的初级抽样单位,其中样本量为初级抽样单位中的居民号码的数量,在每个被选的初级抽样单位中再抽取一个固定数量的居民号码。

正如我们已经看到的,两阶段抽样的使用与有相同样本量的单阶段样本相比,一般会导致精度更低的抽样估计量;即,其设计效应几乎总是大于 1 的。而当人们考虑到使用两阶段抽样方法的经济性及其需要相应增加的样本量超出了其精度的损失的时候,这一方法就是合适的。而使用瓦克斯伯格方法的原因在于其能够产生更高比重的居民号码。在

这一方法下，群中大约2/3的号码为居民号码，这与前面简单方法所产生的1/5有很大差别。

由于缺乏分层变量的信息，在电话抽样中使用分层是非常受限的。美国电讯提供的区域代码/交换中心码组合所构成的抽样框仅仅提供了每次交换的纵向和横向的坐标——一个覆盖一组或一个交换中心码的地理单位——以及每次交换所包含的交换中心码的数量。根据这一信息，样本可以按照地理位置（使用中转地理坐标）以及通过中转的规模（使用被中转覆盖的交换中心码数量作为其规模的指标）来分层。格罗夫斯和卡恩（Groves and Kahn，1979）提供了更多的细节并描述了这一信息在分层中的使用，即用分层因子对一个地区码/交换中心码组合的列表进行排序，然后在这一列表中使用系统抽样方法。

一些电话调查对住户搜集数据，其中访问者需要对指定的受访者或者受访者的集合进行访问。其他调研则对特定的个人搜集数据，而其常常为成人，其中居民号码能够确定一些元素的集合。这一抽样框的问题可以通过随机从中选取一个成人来解决，并且在分析中使用合适的加权方法。一个方法是使用我们在第8章中提到的基什表，但一些研究者认为在调查开始时按照性别和相对年龄进行列表对于访问员来说比较困难并且容易导致较高的据访率。因此，楚德和卡特（Troldahl and Carter，1964）发展出一种技术来避免这一列表，只需要访员搜集该户中合格个体的数量以及合格男性（或女性）的数量。然后访问员可以参照一个表，其中合格个体的数量为一个轴而合格男性的数量为另一个轴，然后从对应的格子中读出被选中的个体编号；这个格子可能是要求

选取“最年长者”。与基什表一样，有很多不同设定的不同版本的表格可以决定选取谁，而对于不同样本，表格往往也不同。使用楚德—卡特方法，我们需要四个表格，每一个都具有相同的频率。布赖恩特(Bryant, 1975)提出了一个修正性的方法来解决样本中男性过少的问题；她的方法对分了其中一个使得女性有更高概率被选取的表格的使用。虽然可能有偏，但是这些替代性的方法在实践中仍被广泛应用(一个实验性的比较，参见 Czaja et al., 1982)。

电话访问也面临着重复元素的抽样框的问题，因为一小部分的家庭有不止一个电话号码。这一问题可以通过从每个抽取的家庭中搜集其电话号码的信息来解决，具体是将与其电话号码数量成反比的因子包含到抽取的家庭或个人的权重中。

第13章

非概率抽样

尽管本文着重介绍概率抽样，但由于非概率抽样也被广泛应用，我们不能对其避而不谈。这一部分讨论了部分类型的非概率抽样，包括被广泛应用的配额抽样（quota sampling）技术。

概率抽样的主要优势在于概率选择机制允许使用统计理论来验证样本统计量的性质。因此人们使用的估计量是有很小甚至没有偏误的，同时也可以得到样本估计量的精度。非概率抽样方法的弱点在于无法发展相应的理论，因此，非概率样本只能根据主观标准评判。另外，即使经验表明非概率方法在过去表现不错，但这并不意味着它以后也会这样。然而，除了这些弱点，不同形式的非概率抽样在实践中被广泛应用，主要是为了节省成本以及操作便利。

一种类型的非概率抽样有多种名称：偶然抽样（haphazard）、便利抽样（convenience）或者意外抽样（accidental sampling）。下面是一些例子：

——某调研的志愿者被试；

——某医生的病人们；

——某学校的孩子们；

——在某街角进行的访问；

——某杂志中的问卷的应答者；

——某希望得到反馈的电视节目中打进电话的观众。

考虑到这些样本潜在的风险，使用它们得到的结果对总体进行统计推断是非常危险的。

另一种非概率抽样被叫做判断抽样(judgment sampling)或者立意抽样(purposive sampling)，抑或专家选择(expert choice)。在此情况下，某一样本是专家选取的，从而使得受访者具有"代表性"。举例来说，一个教育研究者选择某城市的一所学校来获得学校类型的一个界面。在实际中，不同专家很少会同意什么是"具有代表性的"样本，而很多时候这些判断样本(judgment sample)都有一定的主观风险。

随着样本量的增加，人们对判断样本的调查估计量的偏误，或者其他非概率样本估计量的偏误的担心也在增加。因此，应比较一个判断样本中的样本估计量与具有相同样本量的概率抽样的估计量。如果样本量很小，概率样本估计量的方差就会很大，在相对意义上判断样本估计量的偏误就不那么重要了。然而，当样本量增加时，概率样本估计量的方差减少，然而判断样本估计量的偏误则变化不大。这一点表明在样本量很小的时候，使用判断样本是合适的，但是当样本量较大时，则应当使用概率样本。因此，如果一个研究者只能在一个或两个城市中展开一项研究，使用专家选取也许比依赖于完全的随机抽样更好，因为后者很容易导致一个奇怪的样本。然而，如果样本量增加到了 50 个城市，那么则应当考虑一个仔细分层的概率样本了。

第三种类型的非概率抽样被叫做配额抽样。这一方法有很多变种，因其成本低廉、容易管理以及比概率样本更容

易执行的优点被广泛使用。这一方法的精髓在于访问员被分配给了他们应当访问的不同类型人的配额。比如，一个访问员可能被分配给了六个35岁以下男性、五个35岁以上男性、五个被雇用的女性以及八个未被雇用的女性的配额。对访问员在这四个组中分配配额的目的是为了避免(至少是控制)访问员在选取其受访者时过于随意的情形。配额控制可能是相关的，正如上面的例子所给出的，或者也可以是彼此独立的——比如，设定10个男性、13个女性、11个35岁以下以及12个35岁以上的配额。

抽取一个全国性的配额样本的开始步骤与全国性的概率样本一样，通常是用概率抽样实现的。仅仅是在最后选取受访者的阶段，这两种类型的样本才有所不同。对于概率样本，访问员需要采访通过概率机制选取的特定的个人，然而对于配额样本，他们需要完成他们的配额，通常也伴有额外的限制，比如他们应在什么时间打电话以及他们应遵循什么样的路线。配额样本的访问员可能还需要寻找适合其未完成的配额的受访者，具体方式是在被抽取的街区内从固定的起始点开始寻找，并且每个居住地点的访问不能超过一个受访者。

我们还需要评论一下在配额抽样中对每一个住址的受访者不超过一个的控制措施。虽然这一控制能够使得样本在不同住址间分布更广泛，并且避免实地调查中在同一住址进行多个访问的问题，但它会使得较大规模的住址的个人被代表不足(underrepresentation)(Stephenson, 1979)。当然，概率样本也常常仅在每个住址中抽取一个人，但在这种情况下，使用与抽取概率成反比的权重能够修正这一代表不

足的问题。

在配额抽样中,通过这种控制方式形成的配额组通常与层相比,因为两者都代表了从不同组中抽取的样本。尽管二者具有相似性,但我们应当意识到两种类型的分组之间有重要区别,即在群中的元素抽取是通过概率方式进行的,但在配额组中则并非如此。这一区别导致了形成层和配额组的准则是不同的。因为概率抽样避免了抽样偏差,层的选择仅仅需要考虑如何提高调查估计量的精度;正如之前看到的,因为层的内部关于调查变量具有更强的同质性,精度的增加可以通过分层实现。另一方面,使用配额抽样的最重要的准则在于最小化选择偏差。为了达到这一目的,形成相对于调查变量来说内部同质性的配额组是有帮助的,但是首先需要考虑的就是在成员是否能够接受访问方面实现同质性,或者说,形成的不同组能达到组与组之间成员受访的可能性不同。考虑到后面一点,芝加哥大学的国情研究中心使用了四个配额组,这在 20 世纪 60 年代和 20 世纪 70 年代的基于“有配额的概率抽样”的调研中被引用,即,35 岁以下的男性(或 30 岁以下);35 岁以上(或 30 岁以上);受雇用的女性;未受雇用的女性(Sudman, 1966; Stephenson, 1979)。这些特别指定的控制可以对一些难以找到的群体,比如,年轻人以及未受雇用女性的产生较好的代表性。通过使用对配额概率抽样方法在被抽取的界区内进行严格的访问员路线的地理控制,足以提供有较好的种族和经济构成的样本。

在选择了配额组后,访问员的配额就由不同组之间的人口分布方面的数据可得性决定了,而这一信息往往由最近十年的普查得到。这些配额的设定可以或多或少地对所有访

问员一致,可以由不同组之间的人口分布方面的数据可得性决定,或者也可以根据访问员工作的抽样地区的分布特征发生变化。如果根据已经决定的配额得到的数据并不准确(可能是因为配额过时了),配额样本的分布就不会服从组间总体的真实分布。这一情况就与能够对这些不准确进行自我修正(self-correcting)的概率设计形成对比。

有时候,配额抽样会被认为能够避免无应答的问题。然而在实际中,配额样本是将不能或者不愿意被访问的成员替换成其他的受访者。因此,尽管配额样本能够在配额控制下产生要求的分布,但是它对于那些很难联系到的或者不愿意参加访问的人而言,代表性依然不足。因此,相对概率样本而言,它实际上更可能对这些人代表不足,而在概率抽样的条件下,访问员需要对已经制定的样本中的成员进行访问采取坚持不懈的努力。

但除了其弱点,配额抽样在实际中被广泛应用主要基于两个原因。第一是在被抽取的地区选取受访者时不需要使用抽样框。第二是访问员不需要回访来联系特定的受访者。在配额样本下,如果访问员通过电话不能联系到一个合格的成员,该访问员可以简单转到下一家。两个特征使得访问更加简单,并且与概率抽样相比,配额抽样的访问可以进行得更迅速。另一个相关的因素是配额抽样的成本更低。然而,这一成本则依赖于控制的大小:相应的控制越不严格,成本就越低,但另一方面,导致严重的选择偏差的风险也会更大。

第14章

结　语

抽样调查是一个高度专业化和迅速发展的调查领域。目前有众多抽样技术可以选择,但也要注意其中的陷阱。抽样调查的初学者需要在其过程中非常小心,因为调查结果的效用会因为抽样设计中的错误而受较大影响。正因为如此,当从事一项调查的时候,对抽样方面不甚了解的明智的研究者应当咨询一个资深的抽样统计师。

抽样调查的理论和实际方面的文献都非常丰富。由于篇幅所限,本书仅仅提供了这一主题的一个概览,但不能使读者理解所有技术的优劣。希望了解更多的读者可以参考这方面的专著。特别推荐基什(Kish, 1965)、汉森等人(Hansen et al., 1953)以及耶茨(Yates, 1981)的著作,因为它们对抽样的实际应用讨论非常出色,另外,科克伦(Cochran, 1977)、苏哈特姆(Sukhatme and Sukhatme, 1970)以及默西(Murthy, 1967)的著作对于抽样理论的讨论很详细。本书使用的符号和术语与基什(Kish, 1965)基本一致,以便读者在阅读这些书籍时进行参考。戴明(Deming, 1960)的书对重复抽样的广泛应用有很好的介绍,同时还提供了一些实际的建议。在较初级的水平,拉吉(Raj, 1972)、利维和莱姆修(Levy and Lemeshow, 1980)以及苏德曼(Sud-

man，1976)的著作都很有用。斯图尔特(Stuart，1976)用非数学的方法通过一个数字很小的例子介绍了抽样的基本理念，而莫泽和卡尔顿(Moser and Kalton，1971)的书中关于抽样的章节也提供了对这一主题的介绍。

参考文献

Blalock, H.M.(1972) *Social Statistics*, New York: McGraw-Hill.

Bryant, B.E.(1975). "Respondent selection in a time of changing household composition". *Journal of Marketing Research* 12:129—135.

Cochran, W.G.(1977) *Sampling Techniques*. New York: John Wiley.

Czaja, R., Blair, J., & Sebestik, J.P.(1982). "Respondent selection in a telephone survey: A comparison of three techniques". *Journal of Marketing Research* 19:381—385.

Deming, W.E.(1960) *Sample Design in Business Research*. New York: John Wiley.

Dillman, D.A.(1978) *Mail and Telephone Surveys*. New York: John Wiley.

Frankel, M.R.(1971) *Inference from Survey Samples*. Ann Arbor, MI: Institute for Social Research.

Frankel, M.R. and L.R. Frankel.(1977) "Some recent developments in sample survey design". *Journal of Marketing Research* 14:280—293.

Goodman, R., & Kish, L.(1950)"Controlled selection—a technique in probability sampling". *Journal of the American Statistical Association*, 45: 350—372.

Groves, R.M. and R.L.Kahn(1979) *Surveys by Telephone*. New York: Academic.

Groves, R.M., & Lepkowski, J.M.(1982). "Alternative dual frame mixed mode survey designs". Proceedings of the Section on Survey Research Methods, American Statistical Association: 154—159.

Hansen, M.H., W.N.Hurwitz, and W.G.Madow.(1953) *Sample Survey Methods and Theory*, Vols.1 and 2. New York: John Wiley.

Hess, I., D.C.Riedel, and T.B.Fitzpatrick.(1975) *Probability Sampling of Hospitals and Patients*. Ann Arbor, MI: Health Administration Press.

Iversen, G.R., & Norpoth, H.(1976)"Analysis of variance". *Sage University Paper series on Quantitative Applications in the Social Sciences* 07—001. Beverly Hills, CA: Sage.

Kalton, G.(1977). "Practical methods for estimating survey sampling errors". *Bulletin of the International Statistical Institute*, 47, 3: 495—514.

Kalton, G. and Kasprzyk, D.(1982, August)"Imputing for missing survey

responses". *Proceedings of the Section on Survey Research Methods*, American Statistical Association: 22—31.

Kaplan, B.A., & Francis, I.(1979)"A comparison of methods and programs for computing variances of estimators from complex sample surveys". *Proceedings of the Section on Survey Research Methods*, American Statistical Association: 97—100.

Kendall, M.G., & Smith, B.B.(1939) *Tables of Random Sampling Numbers*. Cambridge: Cambridge University Press.

Kish, L.(1965) *Survey Sampling*. New York: John Wiley.

Kish, L.(1962)"Studies of interviewer variance for attitudinal variables". *Journal of the American Statistical Association*, 57:92—115.

Kish, L. and Frankel, M.R.(1974)"Inference from complex samples". *Journal of the Royal Statistical Society*. Series B 36:1—37.

Kish, L., & Frankel, M.R.(1970)"Balanced repeated replications for standard errors". *Journal of the American Statistical Association*, 65: 1071—1094.

Levy, P.S., & Lemeshow, S.(1980) *Sampling for Health Professionals*. Belmont, CA: Lifetime Learning Publications.

McCarthy, P.J.(1966)"Replication: an approach to the analysis of data from complex surveys". Vital and Health Statistics Series 2, No.14. Washington, DC: Government Printing Office.

Moser, C.A., & Kalton, G.(1971) *Survey Methods in Social Investigation*. London: Heinemann.

Murthy, M.N.(1967) *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.

O'Muircheartaigh, C., & Wong, S. T. (1981) "The impact of sampling theory on survey practice: A review". *Bulletin of the International Statistical Institute*, 49(1):465—493.

Raj, D.(1972)*The Design of Sample Surveys*. New York: McGraw-Hill.

Steeh, C. G. (1981) "Trends in nonresponse rates, 1952—1979". *Public Opinion Quarterly*, 45:40—57.

Stephenson, C. B. (1979) "Probability sampling with quotas: An experiment". *Public Opinion Quarterly*, 43:477—496.

Stuart, A.(1976)*Basic Ideas of Scientific Sampling*. London: Griffin.

Sudman, S.(1976)*Applied Sampling*. New York: Academic.

Sudman, S.(1966)"Probability sampling with quotas". *Journal of the A-*

merican Statistical Association, 61:749—771.

Sukhatme, P.V., & Sukhatme, B.V.(1970)*Sampling Theory of Surveys with Applications*. London: Asia Publishing House.

Thornberry, O.T., & Massey, J.T.(1978)"Correcting for undercoverage bias in random digit dialed national health surveys". Proceedings of the Section on Survey Research Methods, American Statistical Association: 224—229.

Troldahl, V.C., & Carter Jr, R.E.(1964)"Random selection of respondents within households in phone surveys". *Journal of Marketing Research* 1:71—76.

U.S.Bureau of the Census.(1978)*The Current Population Survey: Design and Methodology*. Technical Paper 40. Washington, DC: Government Printing Office.

Waksberg, J.(1978)"Sampling methods for random digit dialing". *Journal of the American Statistical Association*, 73:40—46.

Warwick, D.P., & Lininger, C.A.(1975)*The Sample Survey: Theory and Practice*. New York: McGraw-Hill.

Welniak, E.J., & Coder, J.F.(1980)"A measure of the bias in the March CPS earnings imputation system". Proceedings of the section on survey research methods, American Statistical Association:421—425.

Yates, F.(1981)*Sampling Methods for Censuses and Surveys*. New York: Macmillan.

译名对照表

the Kish selection grid	Kish 表选择法
proportionate stratification	按比率分层
Probability Proportional to Estimated Size, PPES	按估计规模大小成比例的概率抽样
Probabilities Proportional to Estimate Size, PPES	按估计容量比例的概率抽样
Probability Proportional to size Sampling, PPS	按规模大小成比例的概率抽样
half-open interval	半开区间
half-sample replication	半样本复制
technique of controlled selection	被控选择技术
ratio estimator	比率估计
fractional interval	比率间隔
ratio mean	比率均值
variability	变异
coefficient of variation	变异系数
convenience	便利抽样
Standard Metropolitan Statistical Areas, SMSAs	标准都市统计区
standard deviation	标准偏差
standard errors	标准误差
supplement sample	补充样本
strata	层
classes	层级
intraclass correlation coefficient	层内相关性系数
paired selection design	成对选择设计
lottery method	抽奖方法
sampling fraction	抽样比率
sampling distribution	抽样分布
sampling interval	抽样间距
sampling frame	抽样框

sample surveys	抽样调查
start-up values	初始值
traditional hot deck method	传统热卡法
minor civil division	次级划分
underreprepresentation	代表不足
March Income Supplement of the Current Population Survey	当前人口普查的辅助调查
Jackknife Repeated Replications，JRR	刀切重复抽样
delta method	德尔塔方法
Equal Probability Selection Methods，EPSEM	等概率抽样
area sampling	地区抽样
unique identification	独特识别
gaps	断裂
multistage sampling	多阶抽样
multiphase sampling	多期抽样
Second Stage Units，SSU	二阶段单位
analysis of variance	方差分析
interviewer variance	访问员变异
interviewer effects	访问员效应
disproportionate stratification	非比例分层
nonsampling errors	非抽样误差
unclustered proportionate stratified sample	非聚类的等比例分层样本
unrestricted sampling	非限制样本
stratification sampling	分层抽样
probability mechanism	概率机制
follow-up	跟进
estimate	估计
interpenetrating sampling	贯穿抽样法
overrepresented	过度代表
callbacks	回访

product-moment correlation coefficient	积矩相关系数
simple random sampling	简单随机抽样
hierarchy	阶层
cross-sectional	截面
net changes	净变化
blanks	空白元素
controlled selection	控制选择
purposive sampling	立意抽样
linking procedure	链接程序
two-phase sampling	两阶段抽样
two-stage sampling	两阶段抽样
list	列表
panel rotation	面板轮换
implicit stratification	模糊分层
ultimate clusters	末级群
target population	目标总体
Neyman allocatio	内曼配置
haphazard	偶然抽样
judgment sampling	判断抽样
judgment sample	判断样本
paired selection design	配对选取
quota sampling	配额抽样
matching	匹配
Balanced Repeated Replications, BRR	平衡重复复制
census tract	普查地段
weights	权重
missing elements	缺失元素
clusters	群
census of population and housing	人口和住房普查
screening interviews	筛选访问
design effect	设计效应
time sampling	时间抽样

poststrata	事后层
poststratification	事后分层
prior stratification	事前分层
dual frame design	双抽样框设计
double sampling	双重抽样
table of random numbers	随机数表
Random-Digit Dialing，RDD	随机数位拨号
Taylor expansion	泰勒展开
imputation	填补方法
survey estimator	调查估计量
survey population	调查总体
foreign elements	外来元素
pseudoreplication techniques	伪重复抽样技术
simple random sampling without replacement	无放回的随机抽样
noncoverage	无覆盖
unbiased	无偏
unrestricted random sampling	无限制随机抽样
nonresponse	无应答
systematic sampling	系统抽样
item nonresponse	项目无应答
mail survey	信件调查
selection equation	选择方程
stratification after selection	选择后分层
selection bias	选择偏差
subpopulation	亚总体
domains of study	研究领域
Primary Sampling Units，PSU	一级抽样单位
accidental sampling	意外抽样
contamination of responses	应答污染
simple random sampling with replacement	有放回的简单随机抽样
Finite Population Correction，FPC	有限总体修正
element	元素

collapsed strata	折叠层法
cluster sampling	整群抽样
normal distribution	正态分布
confidence interval	置信区间
duplicates	重复
replicated sampling	重复抽样
duplicate listings	重复列举
expert choice	专家选择
subclusters	子群
subclass	子群体
self-representing	自代表
self-correcting	自我修正
gross changes	总变化
population	总体
total/unit nonresponse	总无应答
longitudinal survey	纵贯研究